A foundation for reliable spatial proteomics data analysis

Laurent Gatto[1,2], Lisa M. Breckels[1,2], Thomas Burger[3], Daniel J.H. Nightingale[1], Arnoud J. Groen[1], Callum Campbell[1], Claire M. Mulvey[1], Andy Christoforou[1], Myriam Ferro[3], and Kathryn S. Lilley[*1]

[1]Cambridge Centre for Proteomics, Department of Biochemistry, Tennis Court Road, University of Cambridge, CB2 1QR

[2]Computational Proteomics Unit, Department of Biochemistry, Tennis Court Road, University of Cambridge, CB2 1QR

[3]Université Grenoble-Alpes, CEA (iRSTV/BGE), INSERM (U1038), CNRS (FR3425), F-38054 Grenoble, France

May 11, 2014

[*]email: k.s.lilley@bioc.cam.ac.uk

Abbreviations LOPIT: Localisation of Organelle Proteins by Isotope Tagging, PCP: Protein Correlation Profiling, QC: Quality control, ML: Machine learning, SVM: Support vector machine, PCA: Principal component analysis GO: Gene Ontology CC: Cellular compartment ER: Endoplasmic reticulum iTRAQ: Isobaric tags for relative and absolute quantitation TMT: Tandem mass tags

Running title Spatial proteomics data analysis

Abstract

Quantitative mass spectrometry based spatial proteomics involves elaborate, expensive and time consuming experimental procedures and considerable effort is invested in the generation of such data. Multiple research groups have described a variety of approaches to establish high quality proteome-wide datasets. However, data analysis is as critical as data production for reliable and insightful biological interpretation, and no consistent and robust solutions have been offered to the community so far. Here, we introduce the requirements for rigorous spatial proteomics data analysis as well as the statistical machine learning methodologies needed to address them, including supervised and semi-supervised machine learning, clustering and novelty detection. We present freely available software solutions that implement innovative state-of-the-art analysis pipelines and illustrate these tools using several case studies, from multiple organisms, experimental designs, mass spectrometry platforms and quantitation techniques. We also propose sound analysis strategies to identify dynamic changes in sub-cellular localisation by comparing and contrasting data describing different biological conditions. We conclude by discussing future needs and developments in spatial proteomics data analysis.

1 Introduction

The knowledge of a protein's sub-cellular localisation is of paramount biological importance and the reliable high-throughput assessment of localisation, and as a consequence mis-localisation of proteins, is critical to our understanding of cellular biology. Indeed, a protein must be localised to its intended sub-cellular compartment to enable interaction with its binding partners and substrates and thus be functionally active. These sub-cellular niches are many fold and include organelles, which are physically isolated from the rest of the cell by lipid bilayers, as well as macro-molecular complexes of proteins and nucleic acids such as the nucleolus, ribosomes or centrosomes. These various microenvironments represent specialised compartments with unique and dedicated functions (1). It has been shown that there is a significant correlation between disease classes and sub-cellular localisations (2) and it is well established that loss or gain of function of proteins in may diseases can be attributed to protein mis-localisations (3, 4, 5), further highlighting the importance for reliable, high-throughput prediction of protein localisation to underpin cell biology research and to inform the clinical and associated drug discovery communities. Organelle proteomics, also termed more broadly spatial proteomics, is the systematic study of proteins and their assignments to distinct spatial cellular sub-compartments and is a field rapidly growing in importance (6). Current experimental designs and multivariate data analyses techniques permit a researcher to collectively infer and track localisation of thousands of proteins and promises to elucidate the coordinated changes in localisation at the whole proteome level.

Experimental organelle proteomics requires both sophisticated experimental designs in order to obtain accurate datasets and elaborate methodologies to analyse and make sense of the data (6). Various experimental designs have been proposed that merely focus on the identification of proteins in single organelles through biochemical purification (pure fraction cataloguing) to more complex methods that utilise quantitative mass spectrometry to elucidate the broad sub-cellular diversity of cells (fractionation by centrifugation approaches). Techniques that employ the former which focus on single or a limited number of organelles

suffer from two major drawbacks: they may give rise to misleading and/or erroneous associations while not revealing a broader, biologically more meaningful picture, and suffer from substantial contamination from incomplete purification/enrichment.

Techniques that employ the latter type of experimental designs to investigate the full complement of sub-cellular niches were pioneered in 2006 by several groups. Dunkley et al. (7) published the Localisation of Organelle Proteins by Isotope Tagging (LOPIT) technique and Foster et al. (8) described the Protein Correlation Profiling (PCP) using label-free quantitation. These methods enable measurement of steady state protein distributions to provide a more realistic insight into their sub-cellular localisation, while overcoming the requirement to purify organelles of interest and discriminate between genuine organelle residents and contaminants. Briefly, these techniques start with the gentle lysis of the cells and separate the intact and complete cell content using successive differential centrifugation steps or gradient-based ultra-centrifugation, thus obtaining a continuous separation of the complete cell content as a function of its density. Several fractions representing differential sub-cellular enrichment are then collected and their respective protein complements are identified and quantified by high resolution mass spectrometry. The relative protein abundances within the fractions represent unique organelle-specific distributions amongst partially enriched fractions. The resulting datasets are then formatted as a matrix, representing the protein quantitation patterns along the fractions, which is then subsequently submitted to further data analyses. To date, spatial proteomics relies extensively on reliable organelle markers and supervised machine learning (ML) to infer proteome-wide localisation. Pattern recognition techniques and classification algorithms such as support vector machines (9), random forest (10) and many others (11) use marker proteins of known localisation to compare and match the density-related profiles of proteins of unknown localisation. A matching profile permits the assignment of the protein to the specific marker organelle.

Here, we present a set of contemporary methods adapted from the fields of statistics and machine learning, to form a robust framework for spatial proteomics data analysis. We

begin in section 2 with an introduction of the quantitative data structure and the notion of organelle markers, and then proceed to present best practices for data processing, data visualisation, quality control and protein localisation prediction. The ad hoc data structures and methodological advances that we describe have been implemented and further developed using a set of flexible software packages for the `R` (12) programming language, namely `MSnbase` (13) and `pRoloc` (11), available under permissive open source licenses from the Bioconductor (14) project. Methodological aspects, the understanding of the pipeline and its critical parameters are explored through applications to empirical case studies. In section 3, we proceed with a novel application of the computational methods presented, applied to the analysis of protein dual-localisation and dynamic spatial proteomics, also termed comparative organelle proteome profiling (15), where protein localisation is compared and contrasted under different conditions.

2 Material and methods for spatial proteomics data analysis

In this section, we present a detailed description of spatial proteomics data and a set of steps leading to trustworthy results, as summarised in Fig. 1. Guidelines for the interpretation and critical assessment through visualisation are described to guide new and experienced organelle proteomics practitioners towards an in-depth understanding of their data.

2.1 Quantitative data

The data that is generated by the typical spatial proteomics experimental designs can be represented in tabular format with features and fractions along rows and columns respectively (as illustrated in Fig. 2, left). The features generally correspond to proteins or protein groups, although peptides can also be used. A second critical set of information is required for further data analysis, namely organelle markers. These are proteins that are defined as

reliable organelle residents and can be used as reference points to identify new members of that organelle. These marker proteins are generally selected by domain experts and play a central role in the data analysis at many different levels, as highlighted in the next sections.

The nature of the experimental design will characterise the size and the nature of the quantitative data in multiple ways. For example, the technical advantage of multiplexing strategies such as isobaric labelling quantitation using iTRAQ (16) or TMT (17) labelling systems, however directly reduces the resolution that can be measured along the separation dimension, as one is limited by the number of tags that can be employed to quantify fractions. This inherently limits the data cardinality and consequently reduces the discriminative power of sub-cellular niches with similar profiles, as illustrated in Fig. 3. These protein profiles represent well characterised mitochondrial and endoplasmic reticulum residents from two different experiments, namely a TMT 6-plex (red) and a label-free experimental (black) design, that have been fully quantified using 6 and 16 gradient fractions respectively. The quantitation of more fractions using additional independent mass spectrometry acquisitions must however be mitigated by the prevalence of incompletely characterised profiles (described in section 2.3 and Fig. 4), due to missing values which are rife in label free datasets due to the semi-stochastic nature of data dependent acquisition mass spectrometry approaches.

In the following sections, we will use published datasets from Dunkley et al. (7), and Tan et al. (18) using augmented markers sets described in Breckels et al. (19), in order to illustrate our findings. These datasets are also distributed as well documented and computer-friendly data structure (13) in the `pRolocdata` (11) package and can be used as input to the software packages used to run the computational experiments described below. Both datasets above used the LOPIT technique to localise integral and associated membrane proteins.

In Dunkley et al. (7), two independent *Arabidopsis thaliana* callus membranes were prepared and fractionated by using self-generating iodixanol density gradients. For each preparation, two iTRAQ 4-plex tags (16) were used to quantify seven different fractions, with one fraction which was the same in both acquisitions. Firstly, fraction 1 (least dense) was

labelled with reagent 114, fraction 4 with 115, fraction 7 with 116, and fractions 11 and 12 were pooled and labelled with 117. Then, fraction 2 was labelled with iTRAQ reagent 114, fraction 5 with 115, fraction 8 with 116, and fractions 11/12 with 117. The total experiment yielded replicated measurements for 8 quantitation values along the gradient. Labelled peptides were separated using strong cation exchange chromatography and analysed by LC-MS/MS on a QSTAR XL quadrupole-time-of-flight mass spectrometer (Applied Biosystems). Tan et al. (18) collected *Drosophila melanogaster* embryos at 0 – 16 hours. The material was homogenised and centrifuged to collect the supernatant, thus removing cell debris and nuclei. Membrane fractionation was performed on an iodixanol gradient and fractions were quantified using iTRAQ 4-plex isobaric tags (16) as follows: fractions 4/5, 114; fractions 12/13, 115; fraction 19, 116 and fraction 21, 117. Labelled peptides were then separated using strong cation exchange chromatography and analysed by LC-MS/MS on a QSTAR XL quadrupole-time-of-flight mass spectrometer (Applied Biosystems).

2.2 Data visualisation

The experimental data can be directly visualised by plotting the protein intensities, or generally their relative intensities, along the separation dimension. Although very useful, this representation however is limited by the number of proteins or clusters of proteins that can be represented on one figure. It is often essential to be able to visualise the full dataset on one figure, possibly in summarised form, to assess the underlying structure of the data and develop intuition of what can reasonably be achieved in light of the data's properties. When multivariate data is at hand, a popular technique to describe it is to perform dimensionality reduction. This allows one to represent the data in a reduced set of dimensions, generally 2 or 3, instead of the original higher number corresponding to the individual fractions (the columns of our data matrix represented in Fig. 2), while maintaining as much of the initial information as possible. Principal component analysis (PCA) or nonlinear versions thereof, are procedures that transform the original data into a set of orthogonal components that

are ordered according to the amount of variability that they describe (Fig. 2, right). For a well structured dataset, representing it as a projection along the two first components is often an effective means to obtain a simple, yet representative visualisation of the data. The first principal component, representing the most variability in the data, generally equates to the main separation dimension applied to the cell content. In addition, if one considers the amount of variability that is described along each of these first principal components, one can assess how faithfully this two-dimensional representation of the data describes the high dimensional data. Note that relatively minor variances within the data can also be biologically relevant, and lower components may also be informative. While this representation remains a simplification, it is often possible to gain notable understanding of a complex experiment from this single figure.

2.3 Data processing

In this section, we will discuss two essential aspects of data processing, namely imputation of missing values and data normalisation. Although rarely exposed, it is important to acknowledge the detrimental effect of missing values that are so prevalent when combining independent acquisitions. None of the machine learning algorithms that have been applied to organelle proteomics nor any of the contemporary machine learning methods can directly deal with incomplete data; missing values are always explicitly or implicitly imputed before submission to an algorithm. Missing data and the impact of imputation has not been thoroughly addressed in proteomics, let alone in spatial proteomics. Published LOPIT studies (7, 18, 20) excluded proteins that presented missing values across replicated experiments. PCP studies (8, 21) limited the computation of their χ^2 metric to pairs of fractions (defined as the squared deviation of the normalised profile for all peptides divided by the number of data points), thus increasing the bias by lowering the number of data points. Other studies have either explicitly (22) or implicitly (10) applied data imputation without necessarily assessing the biases of the procedure. To illustrate the issues of imputation of missing data,

we have assessed the impact of imputation using the specific goals of spatial proteomics, i.e. the identification of sub-cellular protein clusters and the assignment of proteins. In Fig. 4, we have used data from (7), which provides complete profiles for 16 fractions (some being replicates) for nearly 700 proteins. After random assignment of missing values and data imputation using nearest neighbour imputation (23), we estimated the effect of the imputation method by tracking the shift of the imputed value with respect to the original data points. As expected, we demonstrate an increasing impact of imputation on the data with the number of missing values (Fig. 4, left). In addition, data imputation results in a translation of data towards the centre of the figure, corresponding to less pronounced protein profiles across the gradient (Fig. 4, right). This trend is representative of a loss of signal resulting of imputation, which results in a reduction in the classification power and a bias toward mis-classification to organelles that are characterised by such *average* profiles, such as plasma membrane in the example shown.

Data normalisation is a topic that has been frequently explored in many areas of transcriptomics and proteomics, albeit never in the light of organelle proteomics data. In all subsequent analyses and visualisations we will use relative intensities across the fractionation scheme. When using absolute intensities for visualisation, the absolute component of the signal will overwhelmingly influence the data transformations and eventually hide the relative signal that is of primary interest. All published research tends to divid each intensity by the maximum, or by the sum of intensities in each row of the data matrix. More work on the benefit of the application of more sophisticated technique would be welcomed, in particular when multiple experimental conditions acquired during different runs ought to be compared (section 3.2). Finally, as illustrated in Fig. 5, the accuracy and precision of the underlying quantitation methodology is an essential parameter for optimal cluster resolution (24), and advances in mass spectrometry technologies and quantitation protocols play a crucial role in the production of reliable data.

2.4 Importance of organelle markers

An organelle marker is a protein known to be a resident of a specific sub-cellular niche in the species and condition of interest. From a computational point of view, markers allow the mapping of regions in the multidimensional data space to sub-cellular localisations (Fig. 2). The validity of markers and thus the reliable mapping of biological information to the multivariate data is generally assured by expert curation of the proteins in the data set. The Gene Ontology (GO) (25), and in particular the cellular compartment (CC) namespace are an essential starting point for protein annotation and marker definition. Nevertheless, automatic extraction of GO CC is only a first step and requires additional curation to counter unreliable annotation based on data that is inaccurate or out of context for the biological question under investigation. While proteins with genuine multiple localisations are of particular interest (see below), one must be careful when assessing multiple GO CC terms and distinguish proteins present in more than one subcellular niche (multi-localisation) from changes in localisation under different conditions and incorrect annotation. Annotations often represent a default/consensus state, while the information required to define a reliable marker is specific to the system under study and the conditions of interest. In particular, examples that do not match an annotated localisation for biologically relevant reasons will be of remarkable interest. As such, there is an inevitable trade-off that must be considered in using a very stringent high-confidence list or increasing the number of markers to better characterise the multivariate data. Both aspects are important, as a minimum number of markers is required for further data analysis, while mis-assignment can have a detrimental effect on data analysis (see sections 2.6 and 2.7 on novelty detection and classification) and a good balance can be obtained through careful quality assessment of the markers (next section).

While marker definition is an important step, it is only the initial step of the analysis workflow and is often time consuming. To facilitate the identification of markers, we have curated proteins and mined publicly available datasets (7, 18, 20, 8, 19, 22, 26, 27) to provide

marker sets for *Arabidopsis thaliana* (543 markers, 13 organelles), *Drosophila melanogaster* (220 markers, 11 organelles), *Saccharomyces cerevisiae* (128 markers, 12 organelles), *Gallus gallus* (102 markers, 5 organelles), mouse (648 markers, 21 organelles) and human (507 markers, 16 organelles) as part of our software infrastructure (11) (see details in supplementary file 1). These markers are now available in the `pRoloc` software and can be directly added to quantitative datasets. They are provided as a starting point to generate reliable sets of organelle markers but still need to be verified against any new data in the light of the quantitative data and the study conditions.

2.5 Quality control

The quality of the data is often evaluated using a set of dedicated quality metrics and/or through visualisation. Unsupervised machine learning methods (clustering) that represent the quantitative data without additional external qualitative information is an efficient approach. Our first quality assessment aims at verifying whether the 1^{st} principle behind gradient-based organelle proteomics is met. Based on De Duve's principle (28), we expect that proteins that share the same sub-cellular localisation should co-localise accross a fraction scheme, resulting in well defined structure in the data. We routinely apply the PCA representation described above, without any additional information (symbols colouring depending on external information like organelle residency), to inspect the data and assess if structure can be observed. In a first instance, overlaying markers can be misleading by conferring a false sense of data structure and should be avoided to inspect the data in a complete unsupervised way; the first quality assessment ought to inform on existence of clusters and structure of data prior to the mapping on biologically relevant niches. If no structure is present, even if coherent marker groups can be identified, one should not expect well defined classification boundaries that separate the sub-cellular clusters and thus interpretation of data points located in the continuous cloud of points separating two *clusters* will be challenging. A second assessment laid out by the experimental design can be explored by

overlaying meta-data on the PCA plot, in particular organelle markers. These should match, to some extent, the underlying data structure and explain some, not necessarily all, of the observed protein clusters. An important question arises when marker proteins show substantial deviation from the rest of the group, or more generally when a supposedly well-defined cluster shows a widespread, undefined distribution. Consistent lack of structure/clusters in the data is often indicative of poor separation and undermines all subsequent analysis and interpretation. When individual outliers are detected, it is advised to verify the reliability of the data (identification and quantitation accuracy) and annotation trustworthiness. When any of these can be questioned, the annotation or possibly the protein altogether might be removed from the data. If neither mis-identification nor unreliable quantitation can explain the unexpected position of the marker, it will be the experimenters' responsibility to decide whether to un-label the protein, i.e not consider it as a reliable marker despite anticipated localisation and reliable identification/quantitation, or keep it as is and instruct subsequent algorithms of a possible extended mapping of the organelle to the data. It is however important to note that a marker's localisation can not be (automatically) undone during the data analysis (they represent rigidly imposed constraints that anchor the data space) while an unlabelled protein can be assigned any of the identified localisations. A possible approach could be to un-label the unreliable marker and verify if it eventually gets assigned to the expected localisation. The drawback of a systematic application of this approach is an under-representation of the multidimensional data space: un-labelling markers corresponds to a loss of information, and it is, in the end, up to the expert to define if the information is reliable or not and on what grounds it should or shouldn't be trusted in the light of the data and its quality.

It is important to highlight that it is generally not possible, nor desirable, to identify the complete sub-cellular diversity using markers at this stage. In general, reliable markers can easily be identified for large and well studied niches. The nature of supervised machine learning methods that have been used to date in organelle proteomics studies (see section

2.7) constrain assignments of proteins to a set of marker classes. As such, there is a need for discovery of new data-specific and relevant localisation clusters using a reduced set of highly reliable markers and the underlying structure of our quantitative data.

2.6 Novelty detection

The assignment of proteins to organelles in spatial proteomics traditionally relies on supervised multivariate statistical and machine learning analysis wherein a set of highly curated organelle markers (labelled training data), that belong to a finite set of organelles, is used to map gradient profiles of unknown localisation to subcellular localisations with high accuracy. The application of such methods, however, is often hindered by failure to extract organelle markers that cover the whole subcellular diversity in the data, which leads to prediction errors as protein profiles of unknown localisation can only be associated to organelles that appear in the labelled training data. The extraction of all organelle and organelle-related clusters is a difficult task owing to the limited number of marker proteins that exist in databases and elsewhere, and the time-consuming nature of obtaining reliable markers. To address these issues Breckels et al. (19) developed a novelty detection algorithm that is able to identify sub-cellular groupings such as organelles and protein complexes in spatial proteomics experiments. The algorithm, *phenoDisco*, uses a semi-supervised machine learning schema that employs iterative cluster merging combined with Gaussian Mixture Modelling and outlier detection to identify putative sub-cellular compartments. In a semi-supervised scenario a classifier is learned in the presence of both labelled (i.e. organelle markers) and unlabelled (i.e. proteins of unknown localisation) data. In order to apply Breckels' algorithm one requires an initial set of high quality input organelle markers that cover a minimum of two classes each containing six or more marker protein profiles, and of course some unlabelled data which we wish to mine for new phenotype organelle clusters.

The choice of labelled data (i.e. organelle markers), from which the applied machine learning system will learn is extremely important as it can have significant impact on the

success or failure of the learner. To illustrate this paradigm we investigated the impact that marker choice has upon the application of the *phenoDisco* novelty detection algorithm in mining a *Drosophila melanogaster* dataset which had been produced using the LOPIT technology (18). We considered three sources of markers to use as input for the *phenoDisco* algorithm: (i) a highly and manually curated set from experts in the field (20 ER, 6 Golgi, 14 mitochondrial and 15 plasma membrane markers from our curated marker sets, originally obtained from (18)), (ii) unique Gene Ontology (GO) cellular compartment (CC) annotations assigned a localisation from experimental evidence plus those assigned a unique localisation as inferred from structural sequence or similarity in the GO database, and (iii) only unique GO CC annotations assigned a localisation from experimental evidence in the GO database. Reassuringly, it was found that in case (i) where a small set of manually curated markers were used as input, six out of the seven previously unlabelled phenotype clusters that were found in the *phenoDisco* experiments published in (19) were identified i.e. cluster of proteins that represent two ribosomal subunits (40S and 60S), nucleus, proteasome, lysosome and the cytoskeleton. An additional cluster of cytoplasmic proteins was also identified (phenotype 7, Fig. 6 A, right). Remarkably, we found that the use of organelle marker set (ii) has a detrimental effect on the ability of the algorithm to identify any new organelle clusters and we only are able to identify one new phenotype (Fig. 6 B, right). Examination of the organelle markers in set (ii) showed lack of cluster resolution and an overlap between the Golgi apparatus, plasma membrane and endoplasmic reticulum (ER) (Fig. 6 B, left). We also found a mitochondrial outlier in the dataset which was completely separated from the other mitochondrion markers and located towards to ER cluster. It was found that the inclusion of this one clear outlier forced a negative constraint on the phenotype modelling which resulted in a lack of new phenotypes detected. In attempt to improve marker list (ii) we considered marker set (iii) which included unique GO CC annotations assigned from experimental evidence only. We identified 6, 0, 11 and 15 markers for the ER, Golgi apparatus, mitochondrion and plasma membrane respectively. These presented a minimal overlap with

the curated markers (only 3 for the mitochondrion and plasma membrane). We observed a significant improvement in organelle cluster detection (Fig. 6 C, right). We did however see more noise in the form of a number of smaller phenotypes that lay on the edge of the ER cluster. Interestingly, we also noted that the *phenoDisco* algorithm detected the Golgi as an independent phenotype (phenotype 10, Fig. 6 C, right). No unique Golgi CC markers were retrieved from the GO that were assigned from experimental evidence that could be used as input markers in set (ii) thus it is reassuring that we were still able to retrieve this organelle using novelty detection methods. An important step in application of any novelty detection algorithm is the careful examination of the protein content of any new clusters identified. Curation and examination by experts in the field is an essential step in the discovery analysis pipeline. Using such approaches a researcher is able to mine MS datasets at a deeper level and bring to light interesting sub-cellular compartments for more comprehensive validation for use in a supervised machine learning analysis for robust protein localisation assignment.

2.7 Classification

In machine learning the task of classification falls under the broad area of supervised learning. In supervised learning the aim is to train a classifier to learn a mapping between a set of observed instances and a set of associated external attributes that are being predicted (usually known as the class label or predictor). This set of instances along with their known class labels is typically called the *training data*. Once a classifier has been learned from the training data, the aim is to use this classifier to predict the class labels on data with unknown attributes. All methods to date that have been applied to predict protein localisation have used supervised machine learning.

In terms of protein localisation prediction using data from MS-based organelle proteomics experiments, each training data example consists of a pair of inputs: the actual data, generally represented as a vector of numbers (such as the associated normalised ion intensities along a set of fractions for a given protein) and a class label, representing the membership

to exactly one of multiple possible organelle classes (this is usually referred to as multiclass problem). When there are only two possible classes this is referred to as binary classification. Before one can generate a model on the training data and classify unknown residents one has to take care of properly setting the model parameters. Wrongly set parameters can have adverse effects on the classification performance and success of the learner to the same degree as using inappropriate training examples. An important factor to consider in ones choice of training examples i.e. organelle markers, is how well they represent the multivariate data space i.e. the distribution of proteins over which the system performance will be measured. In general, it has been found that learning is most reliable when the training data follow a distribution similar to that of the examples we wish to classify.

Parameter optimisation can be conducted in a number of ways. One of the most common ways to optimise ones parameters is to use the convention of a *training set* (to model) and a *testing set* (to predict) which are subsets extracted from the labelled training data. Observed and expected classification results can be compared, and then used to assess how well a given model works by getting an estimate of the classifiers ability to achieve a good generalisation i.e. that is given an unknown example predict its class label with high accuracy. A commonly used measure of classifier performance is the macro *F1* score, $F1 = 2\,\frac{precision \times recall}{precision + recall}$, which is the harmonic mean of $precision = \frac{tp}{tp+fp}$ and $recall = \frac{tp}{tp+fn}$, such that $tp = true\ positives$ and $tn = true\ negatives$. This procedure is usually used for a range of possible model parameter values (this is called a grid search), and the best performing set of parameters is then used to construct a model on all markers and predict un-labelled proteins. Estimation of the algorithmic performance can be assessed in many ways, such as via cross-validation. In the `pRoloc` package, algorithmic performance is estimated using stratified 20/80 partitioning, in conjunction with five-fold cross-validation in order to optimise the free parameters via a grid search. This procedure is usually repeated 100 times and then the best parameter(s) are selected upon investigation of associated macro *F1* scores. A high macro *F1* score indicates that the marker proteins in the test dataset are consistently correctly assigned by

the algorithm. Often more than one parameter or set of parameters gives rise to the best generalisation accuracy. As such it is always important to investigate the model parameters and critically assess the best choice. The best choice may not be as simple as the parameter set that gives rise to the highest macro *F1* score and one must be careful to avoid overfitting and to choose parameters wisely.

Once the best parameters have been selected they can then be used to build a classifier from the training data of organelle markers. The classifier will return a classification result for all unlabelled instances in the dataset corresponding to their most likely sub-cellular compartment. In addition, it is possible to extract classification accuracy scores that can inform on the reliability of the assignment. Many supervised machine learning algorithms have been developed, some of the most popular being the support vector machine (SVM), K-nearest neighbour (k-NN), random forest, neural networks and naive Bayes among others. These methods, along with newer state-of-the-art algorithms such as the *Perturbo* (29) classification algorithm, are available in the `pRoloc` package. With the vast number of classification methods available it is often a daunting task to choose which method is best suited to the classification task, however it is not often the choice of algorithm that underpins robust results. In fact, it is widely accepted that it is not algorithm choice that matters but the way in which it is applied and the availability of good training data.

As an example of an application of protein localisation prediction using supervised machine learning, we took the first replicate from Tan et al. (18) and applied a weighted SVM classifier for protein classification. The labelled training data (Fig. 7, left) was constructed from manually curated markers from Tan et al. (18) which were further refined using Breckels et al. (19) phenotype discovery algorithm. Here, using the `pRoloc` package we employed a weighted SVM with a Gaussian kernel to learn a non-linear decision function on the training data to map proteins of unknown localisation to one of the known organelle classes. Class specific weights were used when creating the SVM model which were set to be inversely proportional to the class frequencies to account for class imbalance. On the training data

the two free SVM parameters, cost and sigma, were optimised over 100 rounds of stratified 5-fold cross-validation via a grid search and the best pair of parameters for the classifier were chosen from evaluation of the macro *F1* scores (Fig. 7, middle). The optimised SVM classifier was then used to predict protein localisation on the unlabelled data (Fig. 7, right). The size of the points in 7 (right) reflects the classification probabilities.

3 Multiple conditions and multi-localisation

The previous sections demonstrate a robust protocol to mine and understand the data, verify its annotation, explore it to identify new clusters and classify proteins to sub-cellular niches, relying on state-of-the art algorithms and proven methodologies. Additional complexity arises when multiple conditions need to be considered to elucidate the dynamic nature of protein localisation or multi-localisation of proteins to multiple sub-cellular compartments. To illustrate key concepts and pitfalls of such analyses, we generated a set of controlled localisation changes using the experiments from Dunkley et al. (7) and relevant curated marker proteins. We modelled changes in localisations and moved proteins from one organelle to another, by updating their observed quantitative data along the gradient by new meaningful values inferred from the same dataset.

3.1 Multi-localisation

The protein databases provide multiple localisation annotation per protein in about 60% of human UniProt entries (see supplementary file 2). While a certain number of these annotations are likely to be erroneous or represent specialisation of identical compartments and do not necessarily imply that proteins multi-localise under identical conditions, dual-localisation is an important aspect that needs to be addressed. We used the complete data from Dunkley et al. (7) to create variable mixture of protein relative abundances along the gradient to simulate dual-localisation. Fig. 8 (left) shows such an example where all

fractions of an ER (blue) and Plastid (orange) marker protein profiles have been combined to generate a set of ER/Plastid mixtures ranging from only ER to only Plastid through 90% ER/10% Plastid, 80% ER/20% Plastid, ... 10% ER/90% Plastid intermediates. The resulting mixture profiles are represented as points on the global PCA plot (Fig. 8, right), and are coloured according to their classification using a support vector machine classifier and the procedure described in section 2.7. As can be seen, some of the mixtures *travel* over the plasma membrane cluster, localised between the end points on the PCA plot, and are classified accordingly. The size of the points along the mixture gradient are proportional to the classification probabilities. ER/Plastid mixtures that closely match plasma membrane profiles are classified as plasma membrane residents. A plasma membrane marker protein is represented in yellow on the mixture profiles (Fig. 8, left) to illustrate the relevance of the classification result. Various mixtures for other dual-localisation scenarios have been modelled and lead to identical scenarios, where intermediate mixtures match intermediate organelle profiles (see supplementary file 3).

Despite the PCA plot being a two-dimensional projection of the data, the results of the non-linear classifier are accurately described for this well-resolved data. These simulations indicate that proportional mixtures of two well defined organelle members that mimic dual-localisation of proteins at various proportions are easily confounded with other sub-cellular compartments. From this result we deduce that reliable inference of dual- and more generally multi-localisation requires additional biological information, and can not only rely on unique proteins. In particular, we suggest that known dual-localised examples that form coherent clusters are desired to reliably identify new examples; single evidence proteins can hardly be distinguished from quantitation noise or from membership to intermediate compartments.

3.2 Trans-localisation

To simulate multiple conditions, we took advantage of the availability of biological replicates in Dunkley et al. (7). The two membrane preparations exhibit technical variability that

represent a considerable challenge when investigating genuine changes in localisation, thus making this example a faithful representation of real use cases, while allowing us to set and control protein trans-localisations. We chose seven marker proteins (see supplementary file 3 for details) and imposed changes to different destination organelles. The original relative quantitation values have been replaced by the mean fraction values of all destination marker proteins. The trans-localisations are highlighted by arrows on Fig. 9. Below, we demonstrate important aspects influencing the analysis of dynamic spatial proteomics data, namely data normalisation, the identification of trans-localisation and, in section 3.3 concerted trans-localisations.

Data normalisation

The two replicates display biological as well as substantial technical variability, as illustrated on Fig. 9, left panel. The first and second replicates are represented by circles and diamonds respectively, and the corresponding pairs of proteins are linked by dotted segments. The colours represent all marker proteins for the 9 sub-cellular niches identified for this dataset. While the mass spectrometry processing is becoming more reliable and reproducible, the density separation gradient is a sensitive operation that is executed manually. While Trotter et al. (9) have demonstrated that combining different gradients that separate different sets of organelles from replicated measures from a single condition, achieves a better separation than each gradient taken separately, it is essential to reduce intra-condition variability to highlight differences between conditions. We have transformed the data using the variance stabilisation normalisation (30), a technique that has already been successfully applied to proteomics data (31). The result is represented on the right panel of Fig. 9 and shows a substantial improved overlap of replicate 1 (circles) and 2 (diamonds).

Identifying trans-localisations

We combined two complementary procedures to search for the seven trans-localised proteins. We have employed the machine learning tools described in section 2 and performed a classification analysis on two replicates with the trans-localised proteins. As shown in table 1, the two replicates mostly agree (values along the diagonal). There are however 107 other discrepancies, including the seven anticipated trans-localised proteins, that are assigned the expected localisations in each replicate (see the supplementary file 3).

We next devised a second selection criterion, based on the rational that trans-localising proteins should be characterised by different quantitation profiles along the gradient. For each pairs of proteins in the replicates, we summed the squared differences of the respective fraction $\log_2$ ratios: $\Sigma_{i=1}^{n}(log_2\frac{frac^i_{rep1}}{frac^i_{rep2}})^2$. The distribution of these distances is shown on Fig. 10, left panel. The distances corresponding to the seven trans-localised proteins are shown in blue. Three of these proteins are clear outliers (above red dashed line, corresponding to the largest non-trans-localised protein). If we consider the smallest trans-localised distance (black dashed line), nine non-trans-localised proteins display larger distances. These nine pairs of proteins are highlighted in red on the PCA plot on Fig. 10; 5 pairs show differently coloured starting circles (replicate 1) and ending diamonds (replicate with trans-localisations), representing false positive (see details in supplementary file 3).

While all seven trans-localisation have been classified as expected and displayed considerably higher sums of squared $\log_2$ ratios than most other proteins, a notable proportion of false positives are present among the top hits. Only proteins exhibiting extreme trans-localisation distances between sub-cellular locations that span the full width of the separation capabilities of the design are likely to be reliably identified. Such mixed results are likely to be generalisable due to high variability during cell content separation.

3.3 Concerted trans-localisations

While it is difficult to identify genuine single trans-localising proteins, a biologically relevant scenario could imply a set of proteins exhibiting the same event of trans-localisation, termed concerted trans-localisation. We have modelled such a case by trans-locating 15 mitochondrial proteins (purple) to the plastid cluster (orange) (Fig. 11, left). The new positions have been generated by adding small amounts of normal noise to 15 plastidial residents. While the changes in localisation are correctly classified (see supplementary file 3), the trans-localisation distances are not large enough to be differentiated from the technical and biological variability between replicates, as anticipated from our previous results. However, we predicted that it should be possible to identify the synchronised displacement of the candidates. To do so, we counted the number of trans-localisations between all possible pairs or organelles (Fig. 11, top heatmap). We then normalised these counts be subtracting reciprocal pairs to balance gain and loss of residents (Fig. 11, bottom heatmap). For example, 6 changes are documented from ER membrane to the ribosome and 5 from the ribosome to the ER membrane, resulting in a net change of 1 in favour of the ER membrane. Indeed, gains and losses should compensate each other in case of random fluctuations while consistent movements produce a systematic decrease and increase of recorded events at the origin and destination of the concerted trans-localisations. The normalised number of trans-localisations reported our expected count of Mitochondrion to Plastid movements. We also observe that random net displacements up to 8 can be observed and the appearance of concerted trans-localisation will only be apparent when enough proteins display synchronised behaviour.

4 Discussion

We have described a typical pipeline of organelle proteomics data and clarified some central machine learning concepts applied to such data. While it is essential to understand the prin-

ciples, requirements, weaknesses and strengths underpinning such analyses before confidently interpreting the results, the availability of the right tools is essential. The recent review by Drissi et al. (32) presents an overview of proteomics methods for sub-cellular proteome analysis. In a section about bioinformatics tools for the analysis of organelle proteomics data, they do not mention the existence of any software that will allow the analysis of such data and only refer to the importance of existing protein sub-cellular annotation and the role of GO. It is interesting to look back of past studies and note which methods have been used and also whether, with hindsight improvements could have been made in application of approaches and reporting of data. This is a useful exercise to undertake, especially in such an emerging field as spatial proteomics data analysis. The first applications of large-scale organelle proteomics data analysis were PCP (21, 8), that calculated the χ^2 metric using in-house tools and the LOPIT (7, 20, 18) that applied Partial Least Square Discriminant Analysis using the commercial SIMCA software (Umetrics, Umea, Sweden). Trotter et al. (9) implemented custom `R` code (12) and used the SVM algorithm from the `kernlab` package (33) but no code for others to repeat his state-of-the-art procedure is provided. Others have applied other contemporary machine learning algorithms, including random forests (10), Naive Bayes (22) and neutral networks (34) but did not provide means to apply their analyses to new data. While proteomics data is commonly being disseminated through appropriate repositories, it is not commonplace to provide reproducibility in terms of software and data analysis despite their recognised importance (35).

Here we have attempted to redress access to code and the ability to reproduce data analysis by performing the analysis and creating illustrations using the `R` language and a set of well documented Bioconductor (14) software add-ons specifically developed for quantitative proteomics data. The `MSnbase` package (`http://www.bioconductor.org/packages/release/bioc/html/MSnbase.html`) (13) allows the consistent management and processing of the quantitative data and its associated metadata while the `pRoloc` package (`http://www.bioconductor.org/packages/release/bioc/html/pRoloc.html`) (11) pro-

vides a visualisation and statistical machine learning (including all the algorithms mentioned above as well as novel ones (29)) framework to analyse and interpret spatial proteomics data. The software allows the implementation of a robust and reproducible analysis pipeline and is flexible to accommodate various designs and foster the development of innovative analysis strategies. The software provides extensive documentation and tutorials for a fully reproducible organelle proteomics framework. Finally, `pRoloc` benefits from the Bioconductor infrastructure and its full integration to various online resources, including, among many others, the Gene Ontology (the `GO.db` package (36)), the UniProt database (the `biomaRt` package (37)) and human proteome atlas (38, 39) (the `hpar` package (40)).

In this study we have also sought to develop analysis pipelines that will be useful to dual/multi- and trans-localisation study designs. Such approaches build on robust single condition classification accuracy that relies on good resolution of the sub-cellular space to reduce inter-organelle variability (well defined clusters) and enable reliable organelle assignments. Trans-localisation studies over additional conditions suffer from additional levels of variability that can partially be addressed in multiple ways. First, the use of biological knowledge, including dual-localised or concerted dynamic protein markers can be used to direct the supervised components of the analyses while providing a reliable starting point to uncover genuine signal from noise. Second, reduction of technical variability through adequate normalisation (section 3.2) or multiplexed designs will be of paramount importance. The balance between the number of fractions and the advantage of multiplexing strategies to reduce inter-run variability and missing data discussed in section 2.1 becomes even more critical in multi-condition designs, when relying on 3 (TMT 6-plex) or 4 (iTRAQ 8-plex) fractions per condition makes it challenging to obtain any well resolved clusters in the data. The advent of higher multiplexing solutions, like TMT 10-plex, promises to optimise dynamic designs by combining sufficient resolution and reducing technical variability. Finally, replication can confer more accurate classification in single conditions (9) and will provide an assessment of uncertainty to support the identification of multi- and trans-localisation

events.

5 Conclusion

The path to reliable data analysis results is never written in stone, in particular for complex experimental designs and multivariate data. There are however certain requirements that are always applicable. Visualisation of the complete dataset is essential to describe its major features; in the case of a spatial proteomics experiment, we have highlighted multiple applications of a dimensionality reduction technique like PCA. This is of course a simplification of the complete data, but can often provide a first inkling on the extent of separation and success of classification. It is also important to set basic assumptions about the data, assess the organelle markers in the light of the data structure, describe how it is processed and assess the effects of the treatments it undergoes. Finally, to what extent is the result of the data classification algorithm reliable must be questioned. Trust in the results will be gained by proper usage of algorithms, quality control of the data and verification that basic assumptions on the data, such as appropriate separation of the data, reliable usage of markers, consideration of biologically relevant diversity, and the algorithms in terms of adequate utilisation and parameter selection, are met. The methodology that we have demonstrated brings us a step closer to meeting the requirements of a trustworthy spatial proteomics data analysis.

Acknowledgements

LG, CMM and MF were supported by the European Union 7^{th} Framework Program (PRIME-XS project, grant agreement number 262067). LMB was supported by a BBSRC Tools and Resources Development Fund (Award BB/K00137X/1). TB was supported by the Proteomics French Infrastructure (ProFI, ANR-10-INBS-08). AC was supported by BBSRC grant BB/D526088/1. AJG was supported by BBSRC grant BB/E024777/ and a generous

gift from King Abdullah University for Science and Technology, Saudi Arabia. DJNH was supported by a BBSRC CASE studentship (BB/I016147/1).

References

(1) Dreger, M. (2003). Subcellular proteomics. *Mass Spectrom Rev* 22.1, pp. 27–56.

(2) Park, S., Yang, J. S., Shin, Y. E., Park, J., Jang, S. K., and Kim, S. (2011). Protein localization as a principal feature of the etiology and comorbidity of genetic diseases. *Mol Syst Biol* 7, p. 494.

(3) Luheshi, L. M., Crowther, D. C., and Dobson, C. M. (2008). Protein misfolding and disease: from the test tube to the organism. *Curr Opin Chem Biol* 12.1, pp. 25–31.

(4) Laurila, K. and Vihinen, M. (2009). Prediction of disease-related mutations affecting protein localization. *BMC Genomics* 10, p. 122.

(5) Kau, T. R., Way, J. C., and Silver, P. A. (2004). Nuclear transport and cancer: from mechanism to intervention. *Nat Rev Cancer* 4.2, pp. 106–17.

(6) Gatto, L., Vizcaíno, J. A., Hermjakob, H., Huber, W., and Lilley, K. S. (2010). Organelle proteomics experimental designs and analysis. *Proteomics* 10.22, pp. 3957–69.

(7) Dunkley, T. P. J., Hester, S., Shadforth, I. P., Runions, J., Weimar, T., Hanton, S. L., Griffin, J. L., Bessant, C., Brandizzi, F., Hawes, C., Watson, R. B., Dupree, P., and Lilley, K. S. (2006). Mapping the Arabidopsis organelle proteome. eng. *Proc Natl Acad Sci USA* 103.17, pp. 6518–6523.

(8) Foster, L. J., Hoog, C. L. d., Zhang, Y., Zhang, Y., Xie, X., Mootha, V. K., and Mann, M. (2006). A mammalian organelle map by protein correlation profiling. eng. *Cell* 125.1, pp. 187–199.

(9) Trotter, M. W. B., Sadowski, P. G., Dunkley, T. P. J., Groen, A. J., and Lilley, K. S. (2010). Improved sub-cellular resolution via simultaneous analysis of organelle pro-

teomics data across varied experimental conditions. *PROTEOMICS* 10.23, pp. 4213–4219. ISSN: 1615-9861.

(10) Ohta, S., Bukowski-Wills, J. C., Sanchez-Pulido, L., Alves, F. L., Wood, L., Chen, Z. A., Platani, M., Fischer, L., Hudson, D. F., Ponting, C. P., Fukagawa, T., Earnshaw, W. C., and Rappsilber, J. (2010). The protein composition of mitotic chromosomes determined using multiclassifier combinatorial proteomics. *Cell* 142.5, pp. 810–21.

(11) Gatto, L, Breckels, L. M., Wieczorek, S, Burger, T, and Lilley, K. S. (2014). Mass-spectrometry-based spatial proteomics data analysis using pRoloc and pRolocdata. *Bioinformatics* 30.9, pp. 1322–4.

(12) R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

(13) Gatto, L. and Lilley, K. S. (2012). MSnbase - an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. *Bioinformatics* 28.2, pp. 288–9.

(14) Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H., and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5.10, pp. –80.

(15) Yan, W., Hwang, D., and Aebersold, R. (2008). Quantitative proteomic analysis to profile dynamic changes in the spatial distribution of cellular proteins. *Methods Mol Biol* 432, pp. 389–401.

(16) Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A., and Pappin, D. J. (2004). Multiplexed

protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* 3.12, pp. 1154–1169.

(17) Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., Johnstone, R., Mohammed, A. K. A., and Hamon, C. (2003). Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* 75.8, pp. 1895–904.

(18) Tan, D. J., Dvinge, H., Christoforou, A., Bertone, P., Martinez, A. A., and Lilley, K. S. (2009). Mapping organelle proteins and protein complexes in Drosophila melanogaster. *J Proteome Res* 8.6, pp. 2667–78.

(19) Breckels, L. M., Gatto, L., Christoforou, A., Groen, A. J., Lilley, K. S., and Trotter, M. W. (2013). The effect of organelle discovery upon sub-cellular protein localisation. *J Proteomics* 88, pp. 129–40.

(20) Hall, S. L., Hester, S., Griffin, J. L., Lilley, K. S., and Jackson, A. P. (2009). The organelle proteome of the DT40 lymphocyte cell line. *Mol Cell Proteomics* 8.6, pp. 1295–1305.

(21) Andersen, J. S., Wilkinson, C. J., Mayor, T., Mortensen, P., Nigg, E. A., and Mann, M. (2003). Proteomic characterization of the human centrosome by protein correlation profiling. eng. *Nature* 426.6966, pp. 570–574.

(22) Nikolovski, N., Rubtsov, D., Segura, M. P., Miles, G. P., Stevens, T. J., Dunkley, T. P., Munro, S., Lilley, K. S., and Dupree, P. (2012). Putative glycosyltransferases and other plant golgi apparatus proteins are revealed by LOPIT proteomics. *Plant Physiol* 160.2, pp. 1037–51.

(23) Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics* 17.6, pp. 520–5.

(24) Jakobsen, L., Vanselow, K., Skogs, M., Toyoda, Y., Lundberg, E., Poser, I., Falkenby, L. G., Bennetzen, M., Westendorf, J., Nigg, E. A., Uhlen, M., Hyman, A. A., and

Andersen, J. S. (2011). Novel asymmetrically localizing components of human centrosomes identified by complementary proteomics methods. *EMBO J* 30.8, pp. 1520–35.

(25) Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25.1, pp. 25–9.

(26) Harner, M., Körner, C., Walther, D., Mokranjac, D., Kaesmacher, J., Welsch, U., Griffith, J., Mann, M., Reggiori, F., and Neupert, W. (2011). The mitochondrial contact site complex, a determinant of mitochondrial architecture. *EMBO J* 30.21, pp. 4356–70.

(27) Ferro, M., Brugière, S., Salvi, D., Seigneurin-Berny, D., Court, M., Moyet, L., Ramus, C., Miras, S., Mellal, M., Le Gall, S., Kieffer-Jaquinod, S., Bruley, C., Garin, J., Joyard, J., Masselon, C., and Rolland, N. (2010). AT_CHLORO, a comprehensive chloroplast proteome database with subplastidial localization and curated information on envelope proteins. *Mol Cell Proteomics* 9.6, pp. 1063–84.

(28) De Duve, C. and Beaufay, H. (1981). *A short history of tissue fractionation.*

(29) Courty, N., Burger, T., and Laurent, J. (2011). PerTurbo: A New Classification Algorithm Based on the Spectrum Perturbations of the Laplace-Beltrami Operator. *in the proceedings of ECML/PKDD (1)*. Ed. by D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis. Vol. 6911. Lecture Notes in Computer Science. Springer, pp. 359–374.

(30) Huber, W., Heydebreck, A. von, Sültmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* 18 Suppl 1, S96–104.

(31) Karp, N. A., Huber, W., Sadowski, P. G., Charles, P. D., Hester, S., and Lilley, K. S. (2010). Addressing accuracy and precision issues in iTRAQ quantitation. *Mol. Cell Proteomics* 9.9, pp. 1885–97.

(32) Drissi, R., Dubois, M. L., and Boisvert, F. M. (2013). Proteomics Methods for Subcellular Proteome Analysis. *FEBS J* 280.22, pp. 5626–5634. ISSN: 1742-4658.

(33) Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). kernlab – An S4 Package for Kernel Methods in R. *Journal of Statistical Software* 11.9, pp. 1–20.

(34) Tardif, M., Atteia, A., Specht, M., Cogne, G., Rolland, N., Brugière, S., Hippler, M., Ferro, M., Bruley, C., Peltier, G., Vallon, O., and Cournac, L. (2012). PredAlgo: a new subcellular localization prediction tool dedicated to green algae. *Mol Biol Evol* 29.12, pp. 3625–39.

(35) Aebersold, R. (2011). Editorial: From Data to Results. *Molecular & Cellular Proteomics* 10.11. eprint: `http://www.mcponline.org/content/10/11/E111.014787.full.pdf+html`.

(36) Carlson, M. *GO.db: A set of annotation maps describing the entire Gene Ontology*. R package version 2.10.1.

(37) Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* 21.16, pp. 3439–40.

(38) Uhlén, M., Björling, E., Agaton, C., Szigyarto, C. A. A., Amini, B., Andersen, E., Andersson, A. C., Angelidou, P., Asplund, A., Asplund, C., Berglund, L., Bergström, K., Brumer, H., Cerjan, D., Ekström, M., Elobeid, A., Eriksson, C., Fagerberg, L., Falk, R., Fall, J., Forsberg, M., Björklund, M. G. G., Gumbel, K., Halimi, A., Hallin, I., Hamsten, C., Hansson, M., Hedhammar, M., Hercules, G., Kampf, C., Larsson, K., Lindskog, M., Lodewyckx, W., Lund, J., Lundeberg, J., Magnusson, K., Malm, E., Nilsson, P., Odling, J., Oksvold, P., Olsson, I., Oster, E., Ottosson, J., Paavilainen, L., Persson, A., Rimini, R., Rockberg, J., Runeson, M., Sivertsson, A., Sköllermo, A.,

Steen, J., Stenvall, M., Sterky, F., Strömberg, S., Sundberg, M., Tegel, H., Tourle, S., Wahlund, E., Waldén, A., Wan, J., Wernérus, H., Westberg, J., Wester, K., Wrethagen, U., Xu, L. L. L., Hober, S., and Pontén, F. (Dec. 2005). A human protein atlas for normal and cancer tissues based on antibody proteomics. *Molecular & cellular proteomics : MCP* 4.12, pp. 1920–1932. ISSN: 1535-9476.

(39) Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., Wernerus, H., Björling, L., and Ponten, F. (2010). Towards a knowledge-based Human Protein Atlas. *Nat Biotechnol* 28.12, pp. 1248–50.

(40) Gatto, L. *hpar: Human Protein Atlas in R*. R package version 1.4.0.

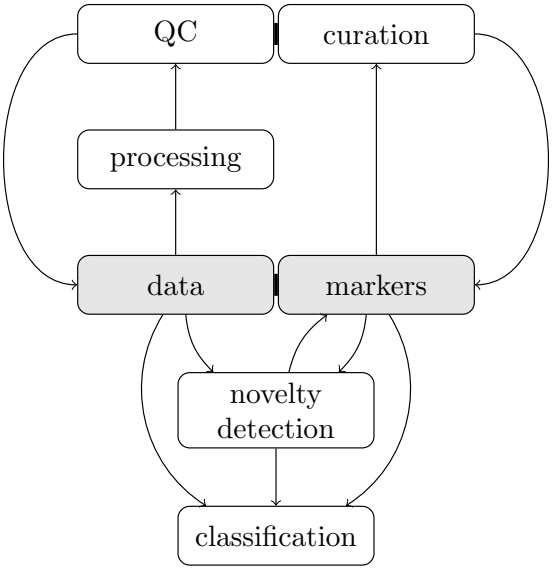

Fig. 1: The steps leading to a sound analysis of spatial proteomics data.

	Fraction_1	Fraction_2	...	Fraction_m
prot_1	$q_{1,1}$	$q_{1,2}$	...	$q_{1,\,m}$
prot_2	$q_{2,1}$	$q_{2,2}$	...	$q_{2,\,m}$
prot_3	$q_{3,1}$	$q_{3,2}$	...	$q_{3,\,m}$
prot_4	$q_{4,1}$	$q_{4,2}$	...	$q_{4,\,m}$
⋮	⋮	⋮	⋮	⋮
prot_i	$q_{i,1}$	$q_{i,2}$	...	$q_{i,\,m}$
⋮	⋮	⋮	⋮	⋮
prot_n	$q_{n,1}$	$q_{n,2}$	...	$q_{n,\,m}$

	markers	
...	*unknown*	...
	$organelle_1$	
	unknown	
	$organelle_2$	
⋮	⋮	⋮
	$organelle_k$	
⋮	⋮	⋮
...	*unknown*	

	Fraction_1	Fraction_2	...	Fraction_m
	...	...	...	...
	⋮	⋮	⋮	⋮
	...	...	...	...

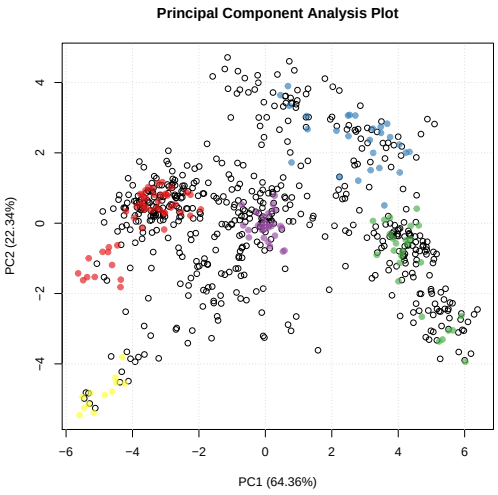

Fig. 2: Left: Representation of a fully described spatial proteomics dataset containing quantitative data for n proteins along m fractions. Each protein is described by additional metadata, in particular definition of known sub-cellular localisation for well known residents. Fractions are also decorated with specific metadata. Right: Summarisation of the quantitative data and annotation of the 'markers' protein metadata using a principal component analysis figure.

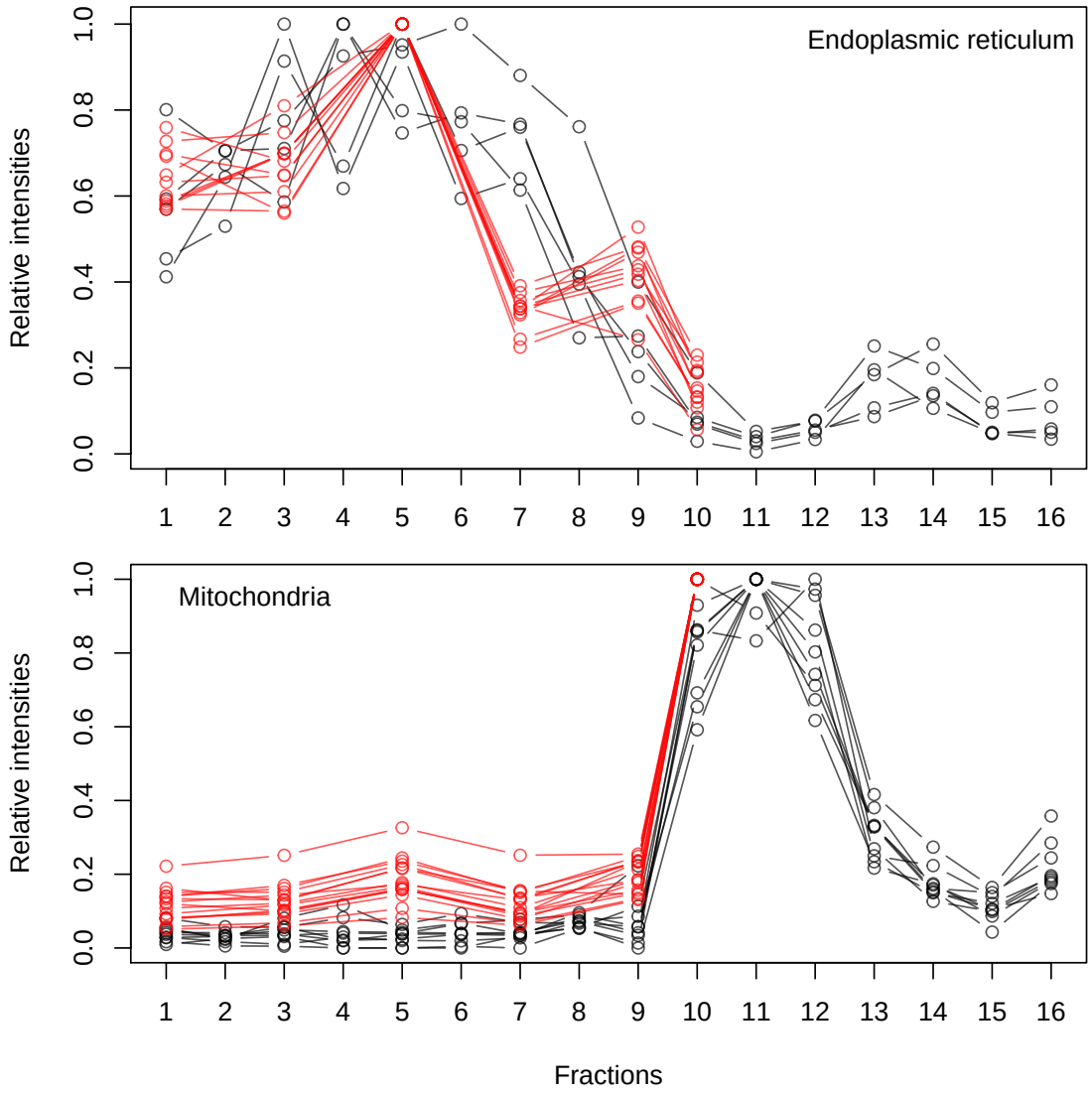

Fig. 3: Effect of the number of fractions on the resolution along the separation gradient for the endoplasmic reticulum (ER) and mitochondria from a TMT experiment using 6 fractions (red) and the equivalent label-free experiment quantified along 16 fractions (black) (unpublished data). The profiles are those of carefully selected *Drosophila* markers that were quantified in all the fractions. 15 and 11 fully characterised mitochondrial and ER markers were identified in the TMT dataset, while using the label-free approach only 9 and 5 complete profiles were identified due to missing values.

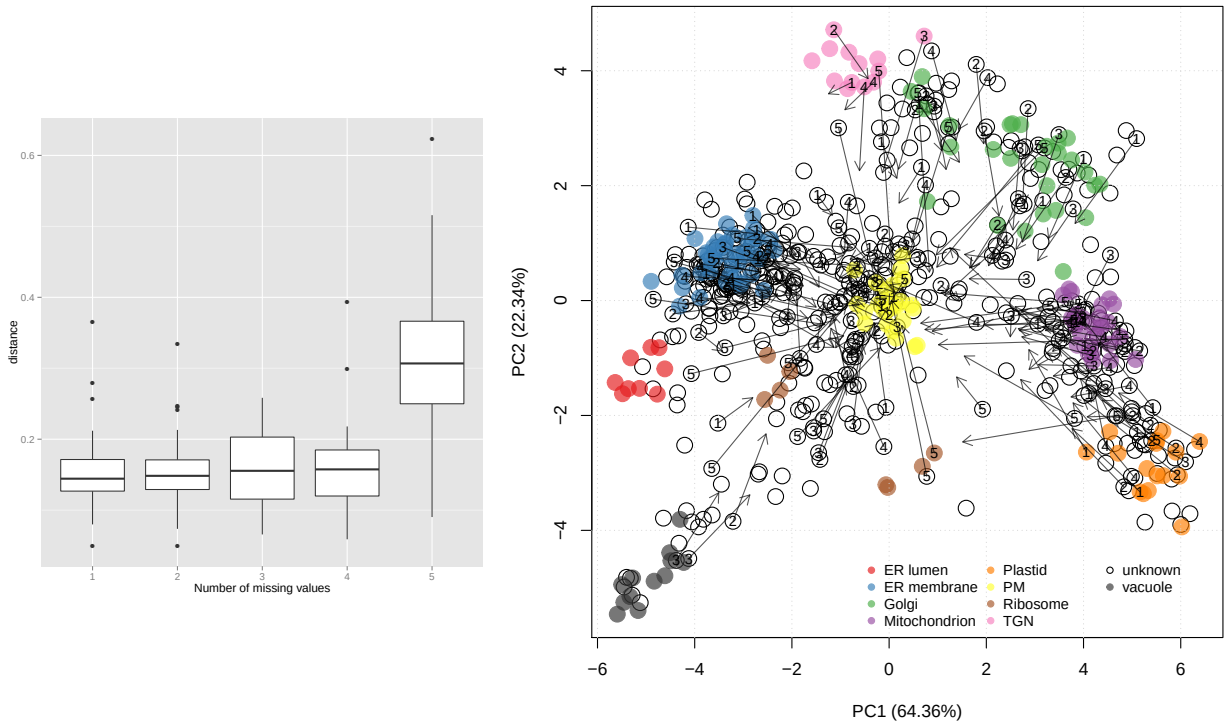

Fig. 4: Assessment of data imputation on cluster resolution and protein organelle assignment. On the left, we illustrate the positive relation between number of imputed values and the displacement of the points before (original values) and after imputation. On the PCA plot on the right, the number of missing values that were imputed are reported in the protein points and the effect on the change in position of the proteins on the PCA plot is highlighed by arrows that show a clear trend toward the origin of the plot.

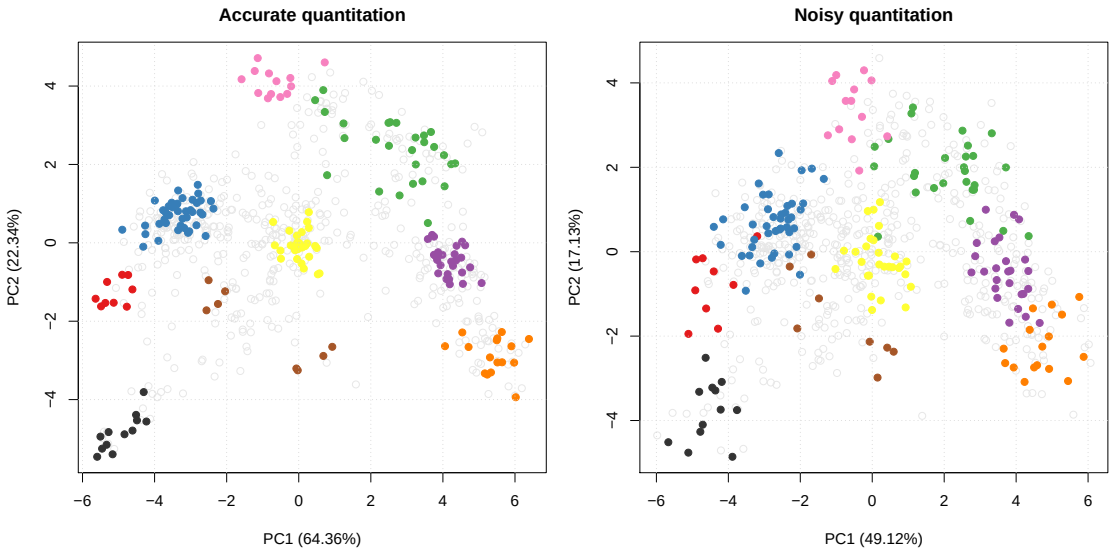

Fig. 5: Effect of noise on sub-cellular cluster resolution. Comparison of original Dunkley et al. (7) data (left) and data to which additional quantitation noise as been added (right). The quantitation noise has been simulated by adding a normal error term (using the mean of the data and $\frac{1}{2}$ standard deviation as parameters) to the quantitation data. While the clusters are still visible and well separated in the noisy data, the original data features much more tight and resolved data with better separation boundaries between groups of interest.

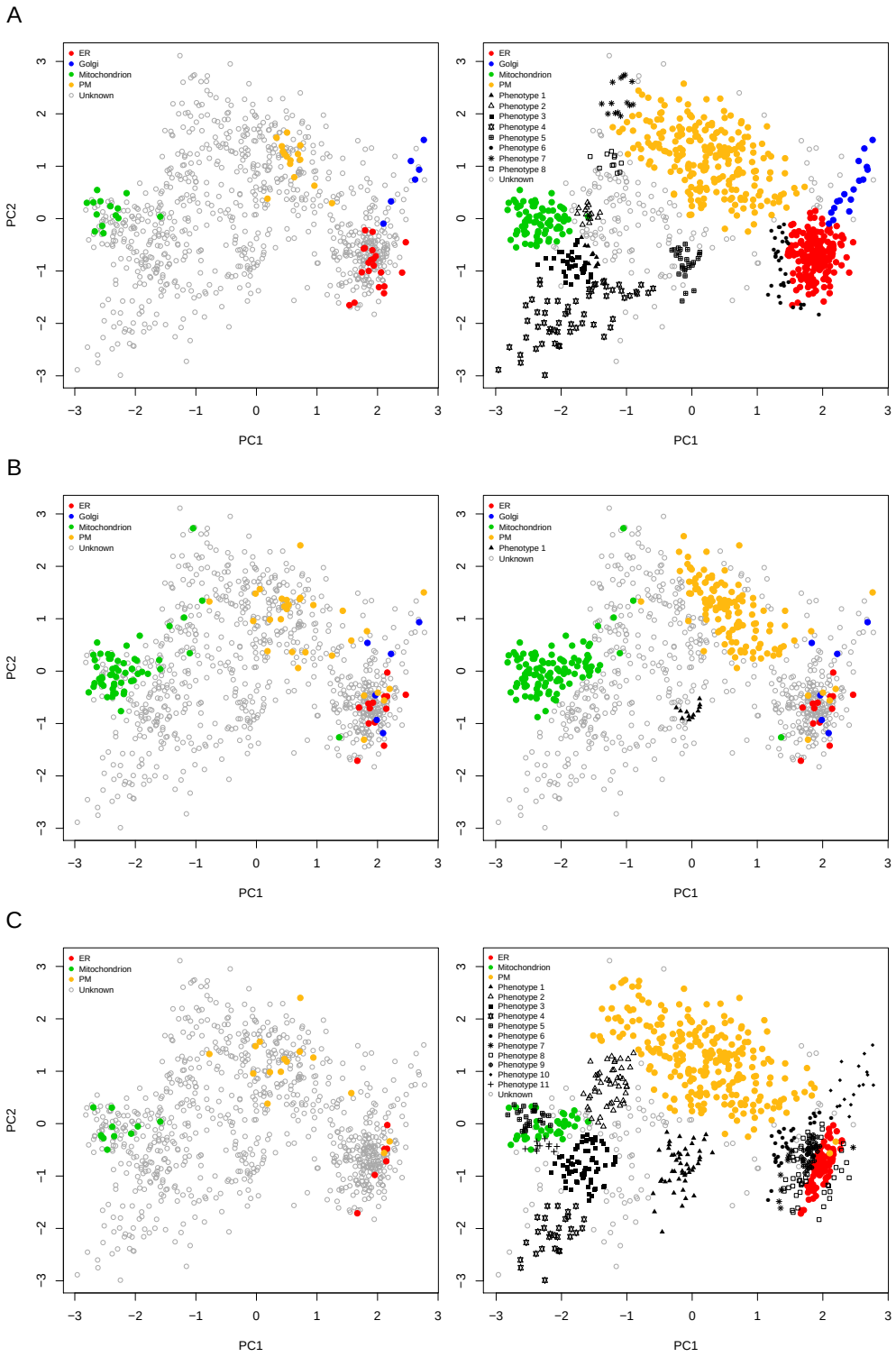

Fig. 6: The effect of different organelle marker sets (left) on the application of the novelty discovery algorithm, *phenoDisco*, (right) in mining a *Drosophila melanogaster* dataset produced using the LOPIT technology (18). (A) A manually curated set from experts in the field. (B) Unique GO CC annotations from experimental evidence or computational predictions. (C) Unique GO CC annotations from experimental evidence.

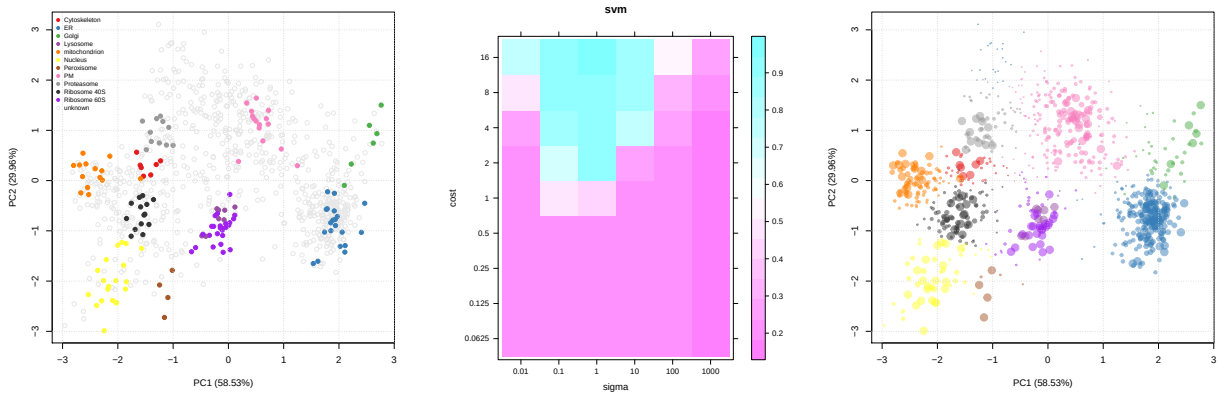

Fig. 7: Application of the support vector machine classifier (SVM) to the Tan et al. (18) data. Left: Augmented dataset after novelty detection. Middle: Grid search for the SVM parameters cost and sigma, highlighting optimal pairs of parameters. Right: Application of the SVM classifier. The size of the points reflects the classification probabilities.

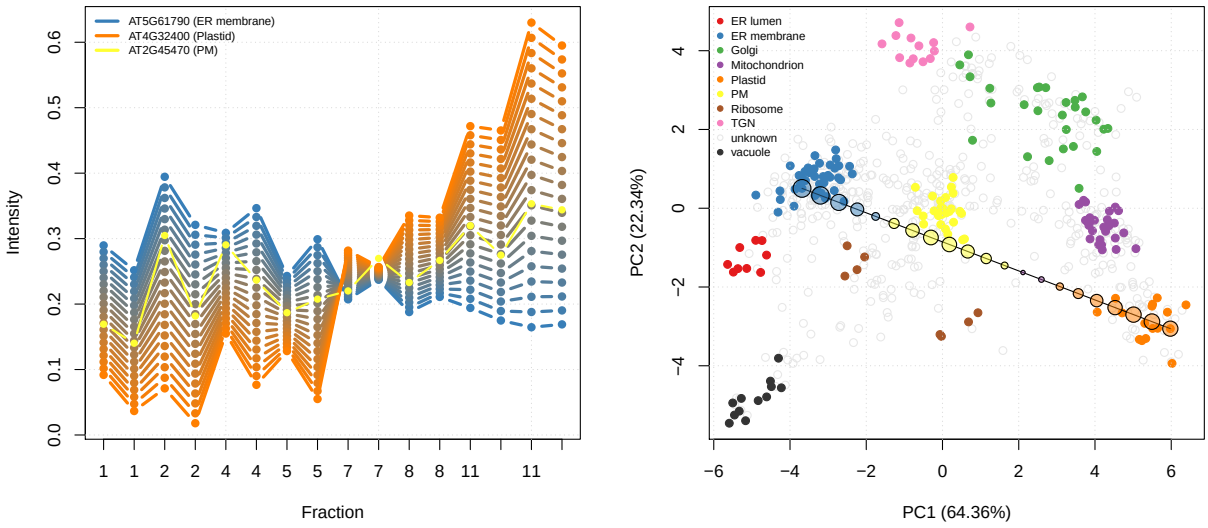

Fig. 8: Applications of the SVM classification algorithm to identify dual-localisation patterns. Left: ER (blue)/Plastid (orange) relative quantitation mixture. A plasma membrane marker protein is shown in yellow. Right: position of the respective ER/Plastid mixtures on the PCA plot and their respective colour-coded classifications.

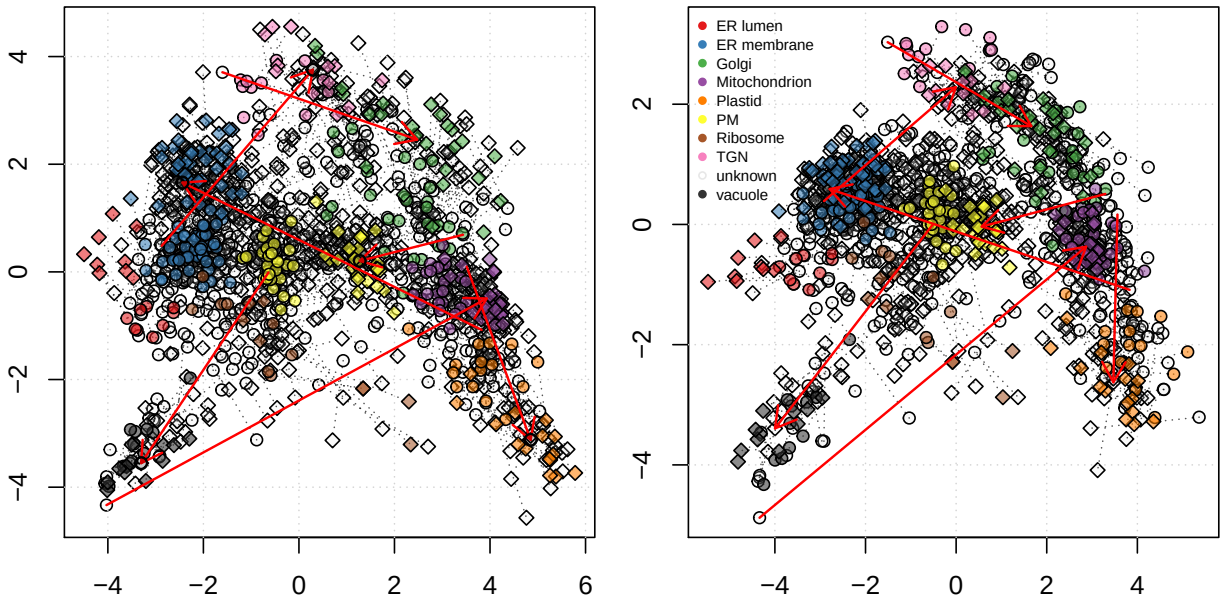

Fig. 9: Application of the variance stabilisation normalisation of two replicates (circles and diamonds) of our test data. The solid red arrows indicate the trans-localised proteins. On the left, original data, showing substantial differences between replicate 1 (circles) and 2 (diamonds). The colours represent all markers for the 9 sub-cellular localisations. After application of the normalisation procedure, on the right, we obtain considerably better overlap between the replicates.

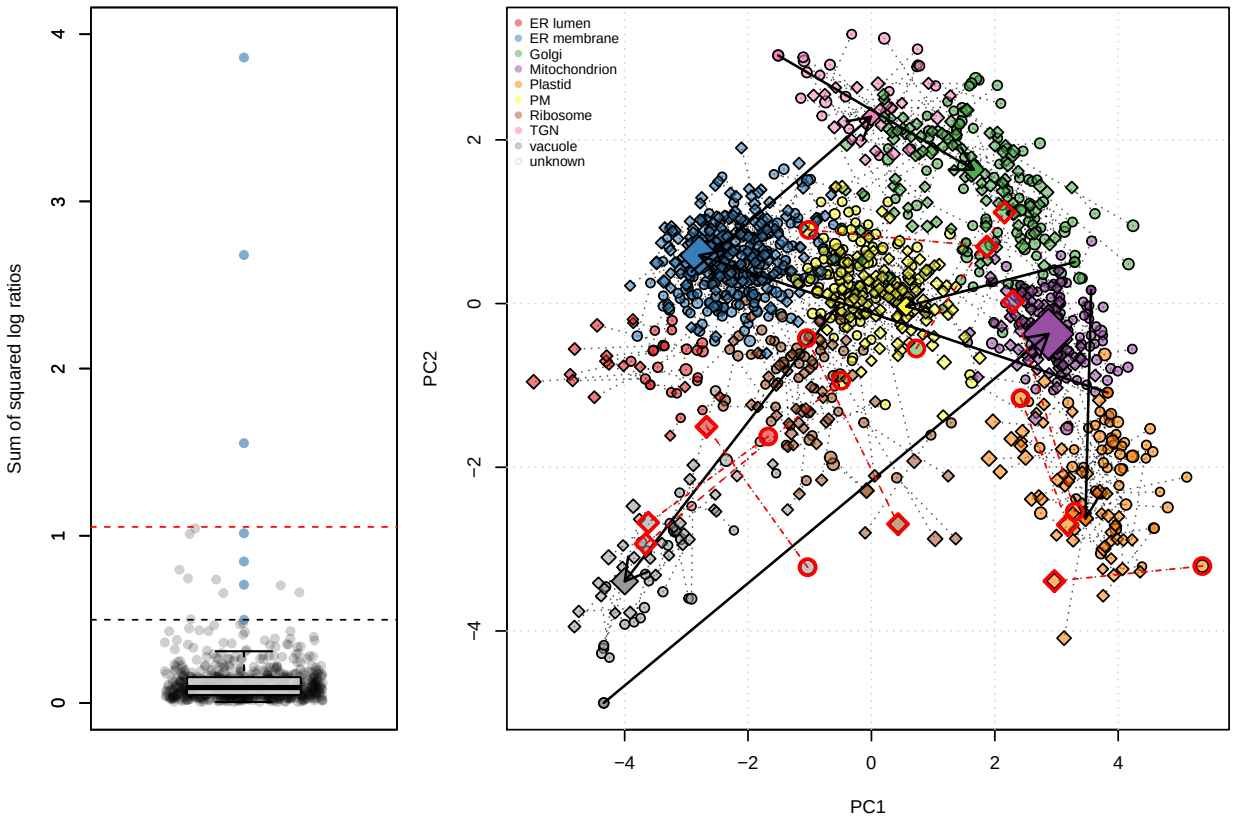

Fig. 10: Identification of changes in localisation. Left: distribution of the summed squared $\log_2$ ratios between fraction of the two conditions of interest. Blue points represent genuine trans-localisations. The red and black dashed lines represent the largest non-trans-localised value and the smallest genuine trans-localised protein. Right: PCA plot illustrating effects of technical variability and trans-localisations. Non-trans-localised pairs with a sum of squared $\log_2$ ratios higher than genuine changes are highlighted in red and genuine trans-localisations are represented by thick arrows.

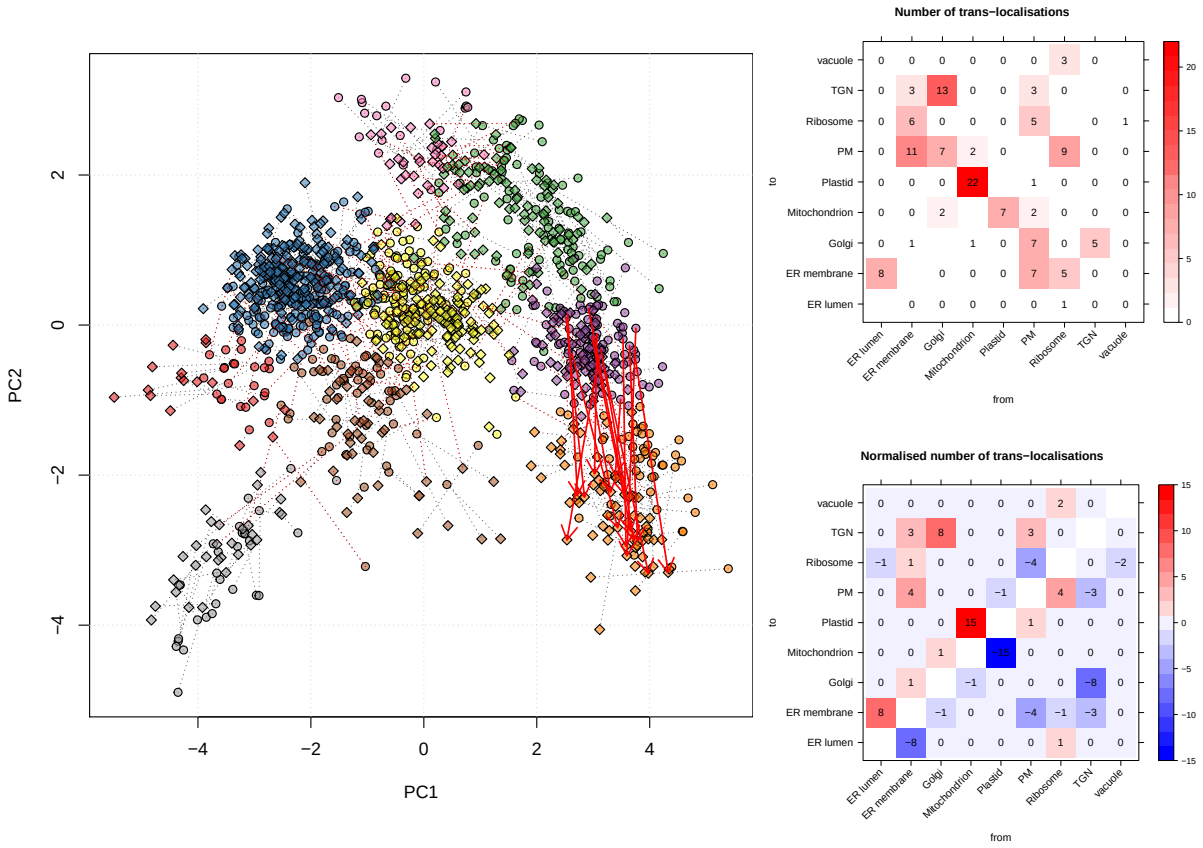

Fig. 11: A scenario of concerted trans-localisation involving 15 proteins moving from the mitochondrial organelle (purple) to the plastid cluster (orange). Random trans-localisations are connected by red dotted segments. The heatmaps on the right show the absolute (top) and normalised (bottom) number of observed trans-localisation effects. Concerted trans-localisation effects are characterised by high net gains and reciprocal losses of displacements.

	ER lum	ER mb	GO	MT	PT	PM	Ribo	TGN	VA
ER lum	16	6	0	0	0	0	1	0	1
ER mb	0	175	3	0	0	9	5	1	0
GO	0	1	81	1	0	6	0	10	0
MT	0	0	0	84	9	2	0	0	0
PT	0	1	1	4	47	0	0	0	0
PM	0	4	5	4	0	98	5	1	2
Ribo	0	4	0	0	0	11	38	0	1
TGN	0	0	7	0	0	0	0	14	0
VA	1	0	0	1	0	0	0	0	29

Table 1: Comparison of the classification results of two replicated experiments including simulated trans-localisations from Dunkley et al. (7). Values along the diagonal correspond to identical outcomes while values in the upper of lower parts of the contingency table represent differences, seven of which are expected based on the imposed sub-cellular changes.