JOURNAL OF PROTEOMICS 88 (2013) 129–140

Available online at www.sciencedirect.com

SciVerse ScienceDirect

www.elsevier.com/locate/jprot

The effect of organelle discovery upon sub-cellular protein localisation☆

L.M. Breckels[a], L. Gatto[a], A. Christoforou[a], A.J. Groen[a], K.S. Lilley[a], M.W.B. Trotter[b,*]

[a]Cambridge Centre for Proteomics, Department of Biochemistry, University of Cambridge, CB2 1QR, UK
[b]Celgene Institute for Translational Research Europe (CITRE), Sevilla 41092, Spain

ARTICLE INFO

Available online 21 March 2013

Keywords:
Organelle
Protein
Assignment
Machine-learning
Prediction
Semi-supervised

ABSTRACT

Prediction of protein sub-cellular localisation by employing quantitative mass spectrometry experiments is an expanding field. Several methods have led to the assignment of proteins to specific subcellular localisations by partial separation of organelles across a fractionation scheme coupled with computational analysis.

Methods developed to analyse organelle data have largely employed supervised machine learning algorithms to map unannotated abundance profiles to known protein–organelle associations. Such approaches are likely to make association errors if organelle-related groupings present in experimental output are not included in data used to create a protein–organelle classifier. Currently, there is no automated way to detect organelle-specific clusters within such datasets.

In order to address the above issues we adapted a phenotype discovery algorithm, originally created to filter image-based output for RNAi screens, to identify putative subcellular groupings in organelle proteomics experiments. We were able to mine datasets to a deeper level and extract interesting phenotype clusters for more comprehensive evaluation in an unbiased fashion upon application of this approach. Organelle-related protein clusters were identified beyond those sufficiently annotated for use as training data. Furthermore, we propose avenues for the incorporation of observations made into general practice for the classification of protein–organelle membership from quantitative MS experiments.

Biological significance

Protein sub-cellular localisation plays an important role in molecular interactions, signalling and transport mechanisms. The prediction of protein localisation by quantitative mass-spectrometry (MS) proteomics is a growing field and an important endeavour in improving protein annotation. Several such approaches use gradient-based separation of cellular organelle content to measure relative protein abundance across distinct gradient fractions. The distribution profiles are commonly mapped in silico to known protein–organelle associations via supervised machine learning algorithms, to create classifiers that associate unannotated proteins to specific organelles. These strategies are prone to error, however, if organelle-related groupings present in experimental output are not represented, for example owing to the lack of existing annotation, when creating the protein–organelle mapping. Here, the application of a phenotype discovery approach to LOPIT gradient-based MS data identifies candidate organelle

☆ This article is part of a Special Issue entitled: New Horizons and Applications for Proteomics [EuPA 2012].

* *Corresponding author at:* Celgene Institute for Translational Research Europe (CITRE), Centro de Empresas Pabellon de Italia, C/Isaac Newton, 4, Sevilla 41092, Spain. Tel.: +34 955 001705; fax: +34 955 001785.
E-mail address: mtrotter@celgene.com (M.W.B. Trotter).

1874-3919/$ – see front matter
http://dx.doi.org/10.1016/j.jprot.2013.02.019

phenotypes for further evaluation in an unbiased fashion. Software implementation and usage guidelines are provided for application to wider protein–organelle association experiments. In the wider context, semi-supervised organelle discovery is discussed as a paradigm with which to generate new protein annotations from MS-based organelle proteomics experiments.
This article is part of a Special Issue entitled: New Horizons and Applications for Proteomics [EuPA 2012].

1. Introduction

Organelle proteomics, the systematic study of proteins and their assignments to organelles, is a field of rapidly growing importance [1]. The determination of a protein's location is desirable to biologists for two reasons. First, it can assist elucidation of a protein's role within the cell, as proteins are spatially organised according to their function and specificity of their molecular interactions [2]. Second, it refines knowledge of cellular processes by pinpointing certain molecular functions to specific organelles [3]. Reliable high-throughput prediction of protein sub-cellular localisation is crucial to both underpin cell biological studies and to inform the medical and associated drug discovery communities. Many wet lab and in silico methods have been developed and applied in an attempt to characterise the protein complement of organelles. Experimental mass spectrometry (MS) based approaches to protein–organelle association are a recent development, but the computational determination of protein localisation is an established bioinformatics challenge, and many methods have been developed to predict protein localisation from amino acid sequence [4–7].

Traditional low-throughput experimental methods involve tagging individual proteins followed by imaging. For example, through the use of fusion proteins such as green fluorescent protein-tagged constructs can be used to determine protein sub-cellular localisation by confocal microscopy [8]. Other approaches involve using immunohistochemistry, involving the use of high 'fluorophore coupled' (or conjugated) antibodies to reveal the sub-cellular location(s) of target proteins [9]. Several MS-based organelle proteomic approaches have been developed for the identification of organelle residents which require sophisticated experimental designs and data analyses in order to obtain accurate datasets [1]. Some methods employ the use of purified organelle fractions, but the total purification of organelles in sufficient quantities for characterisation is in practice exceedingly challenging. For example, components of the secretory pathway are difficult to purify due to the similar physical properties of their membranes. Other organelles, for example the nucleus, mitochondria, and chloroplast can be more easily enriched due to their unique physical properties, but their enrichment may still result in lower levels of contamination by other sub-cellular species unless the enrichment schema is carefully designed and evaluated, and appropriate controls are performed to distinguish true organelle residents from residual contaminants. A further complexity is that many proteins are present in more than one sub-cellular location and so the use of purified organelle fractions does not necessarily reveal what is happening at the cellular level. Furthermore, many proteins traffic from one destination to the other and therefore can reside during transit in organelles where they do have a function.

Several high-throughput quantitative strategies have been developed to overcome the requirement to purify organelles of interest, and instead discriminate between genuine organelle residents and contaminants by determining the unique distribution patterns of known organelle members amongst partially enriched fractions generated by various separation technologies. Some methods examine the enrichment of certain organelle proteins within a small number of highly refined fractions [10–12] whereas others involve more elaborate experimental designs to determine global distribution patterns of proteins in many sub-cellular organelles [13–15]. Localisation of Organelle Proteins by Isotope Tagging (LOPIT) [13,16], employs isobaric tagging coupled with tandem mass spectrometry (MS/MS) [17,18] to capture the distribution of organelle proteins within fractions taken from density gradients. Protein correlation profiling (PCP) is a similar approach which uses label-free quantitation [14,15], and was more recently coupled with SILAC quantitation (PCP–SILAC) [19,20]. Both experiments output a vector (protein profile) of relative abundance measurements that approximates protein distribution along the gradient. Since proteins that belong to the same organelle will co-fractionate within the density gradient [21], given sufficient resolution the distribution pattern of proteins with unknown localisations can be mapped to those of known organelle marker residents.

To date, protein profiles from gradient based data have largely been assigned to organelles using straight-forward statistical analyses methods such as PLS-DA (e.g. [13,16]) and the χ^2 measure (e.g. [14,22]). Supervised machine learning methods, such as support vector machines (SVMs) [23,24], have also been applied to associate protein profiles to organelles with high estimated generalisation accuracy [25], and a recent study by Zhang et al. applied a fuzzy k-Nearest Neighbour algorithm to PCP data to provide a sub-module of their prediction system [26]. In these supervised approaches, a subset of training data, annotated according to known organelle association from public databases, is used to map gradient profiles to sub-cellular locations. The aim is to obtain a mapping that subsequently associates further profiles to the organelles described, thereby annotating them, with high accuracy. Such approaches are highly likely to make association errors, however, if organelle-related groupings present in experimental output are not included in data used to create the protein–organelle classifier. The extraction of all organelle-related groupings is a difficult task owing to (i) a limited number of proteins with known organelle membership; and (ii) the time-consuming nature of obtaining reliable protein annotations from databases. Furthermore, the fact that proteins may be present simultaneously in several organelles makes the learning task more complex. Several approaches towards tackling simultaneous multi-organelle classification have been reported (e.g. [27–32]),

which include the construction of probabilistic classifiers [33–36], capable of expressing the likelihood of observing a protein in more than one organelle. When used on a discrete-labelled training data, such approaches are balanced between reflecting a potential biological truth, i.e. distribution of a protein across multiple organelles, and simply reflecting classifier uncertainty when trained with insufficient information, i.e. training data labelled according to known protein concentrations across multiple organelles. Moreover, the consideration of describing organelle occupancy via classification probabilities is again severely hindered unless all potential organelle groupings resolved along a separation gradient are reflected, at least in part, by the training data.

Here, we present an adaptation of a published phenotype discovery algorithm [37], originally intended to filter image-based output for RNAi screens, which is applied to identify putative organelle groupings in gradient-based MS datasets. The empirical LOPIT data-sets employed were produced from *Arabidopsis* callus [13], *Drosophila* embryos [38], and a HEK293T human cell line (Christoforou et al. unpublished). The algorithm is assessed as to its ability to re-discover well-annotated phenotype clusters when systematically deleted from the data, and is subsequently applied to identify new phenotypic organelle clusters, which are evaluated by manually querying online databases and the literature for biological interpretation. The results obtained demonstrate a need for new approaches to tackle protein–organelle prediction in gradient-based MS data to permit the assignment of proteins to organelles outside of the classes present during classifier creation. The suitability of an alternative learning framework is apparent, which opens avenues for further work.

2. Methods

2.1. Datasets

Three datasets, from studies on non-photosynthetic *Arabidopsis thaliana* callus [13], *Drosophila* embryos [38] and Human embryonic kidney fibroblast (HEK293T) cells (unpublished) were collected using the standard LOPIT approach as described by Sadowski et al. [16]. In the LOPIT protocol fractions from a self-generating iodixanol density gradient are collected and a set of enriched fractions are then digested and labelled separately with iTRAQ reagents, pooled and the relative abundance of the peptides in the different fractions is measured by tandem MS. The number of measurements obtained per gradient occupancy profile (which comprises of a set of isotope abundance measurements) is thus dependent on the iTRAQ reagents and LOPIT methodology used.

The *Arabidopsis* dataset was collected using the duplex method which employs dual use of four isotopes across eight fractions and thus yielding 8 values per protein profiles. The aim of this experiment was to resolve Golgi membrane proteins from other organelles. Gradient-based separation was used to facilitate this including separating as much nuclear material as possible during a pre-centrifugation step and carbonate washing of membrane fractions to remove peripherally associated proteins to maximise the likelihood of assaying less abundant integral membrane proteins from organelles involved in the secretory pathway.

The aim of the *Drosophila* experiment was to apply LOPIT to an organism with heterogeneous cell types. Tan et al. were particularly interested in capturing plasma membrane (personal communication). There was a pre-centrifugation step to deplete nuclei, but no carbonate washing, thus peripheral and luminal proteins were not removed. In this experiment four isotopes across four distinct fractions were implemented and thus yield four measurements (features) per protein profile.

The Human dataset was a proof-of-concept for the use of LOPIT with adherent mammalian cell culture, and employed the use of 8-plex iTRAQ reagents, thus returning eight values per protein profile within a single labelling experiment. As in the LOPIT experiments in *Arabidopsis* and *Drosophila* the aim was to resolve the multiple subcellular niches of post-nuclear membranes, and also the soluble cytosolic protein pool. Nuclei were discarded at an early stage in fractionation scheme as previously described, and membranes were not carbonate washed in order to retain peripheral membrane and luminal proteins for analysis.

In all experiments the existing labelled protein profiles and newly identified phenotype clusters were annotated to known organelles and protein complexes, as evidenced from localisation annotation within the literature [13,38] and reviewed annotation from the UniProt Knowledgebase (http://www.uniprot.org/help/uniprotkb) [39], FlyBase (http://flybase.org/) [40], TAIR (http://www.arabidopsis.org/) [41,42] and PANTHER (http://www.pantherdb.org/) [43] databases. The organelle classes as originally published were used as training classes and the distribution of these markers are displayed in Table 1.

Table 1 – Markers of protein sub-cellular localisations from LOPIT studies on *Arabidopsis* callus [13], *Drosophila* embryos [38] and Human embryonic kidney fibroblast (HEK293T) cells (unpublished).

Arabidopsis callus						
Organelle	Endoplasmic reticulum	Golgi apparatus	Mitochondria/plastids	Plasma membrane	Vacuole	Unknown	Total
Proteins	49	27	26	28	12	547	689

Drosophila embryos [a]				
Organelle	ER/golgi apparatus	Mitochondria	Plasma membrane	Unknown	Total
Proteins	235	74	180	399	888

HEK293T (unpublished)						
Organelle	Endoplasmic reticulum	Golgi apparatus	Mitochondria	Plasma membrane	Unknown	Total
Proteins	82	28	151	61	1049	1371

[a] Markers as annotated from the results from Tan et al.

2.2. Phenotype discovery

2.2.1. Notation and feature extraction

MSnbase [44] is an open source Bioconductor [45] package for the R statistical programming environment (http://www.r-project.org) that provides a framework for the analysis of quantitative proteomics experiments. Using *MSnbase*, all datasets were imported into R and converted to "*MSnSet*" instances. The *MSnSet* class is a computational representation of the data that allows storage and easy manipulation of quantitative expression data and relevant metadata for MS proteomics experiments. In addition, it guarantees validity of the data throughout the various processing and manipulation steps of the analysis pipeline.

Individual datasets consist of multivariate protein profiles which are annotated to either (i) a single known organelle (labelled data), or (ii) have unknown localisation (unlabelled data) i.e. $D = (D_L, D_U)$. Labelled examples are represented by $D_L = (\mathbf{x}_l, y_l)_{l=1,\ldots,|D_l|}$ where each l^{th} protein is described by a vector of p features such that $\mathbf{x}_l \in \mathbb{R}^p$ and is annotated to one of $y_l \in m = \{1,\ldots,K_0\}$ organelle classes (phenotypes). The dataset in the m^{th} organelle class with N_m indicating the number of proteins for the m^{th} organelle class is given by $g_m = (\mathbf{x}_i, m)_{i=1,\ldots,N_m}$ and the labelled dataset of all available proteins over the m different organelle classes is represented by $D_L = \bigcup_{m=1}^{K_0} g_m : \forall m, n \in \{1,2,\ldots,K_0\}, g_m \cap g_n = \varnothing$. Unlabelled examples are represented by $D_U = (\mathbf{x}_u)_{u=1,\ldots,|D_U|} : \mathbf{x}_u \in \mathbb{R}^p$.

To facilitate modelling, principal component analysis (PCA) was applied to each dataset before algorithm application. The two principal components that described the greatest variance within the data were chosen as features, thus reducing the dataset cardinality to $p = 2$. Given a multivariate dataset, PCA transforms the original high dimensional data into a set of linearly uncorrelated variables (principal components) such that the first principal component has the largest possible variance to account for as much variability in the data as possible and each succeeding component in turn has the highest variance possible under the constraint that it be orthogonal to the preceding components.

2.2.2. The phenotype discovery algorithm

A previously published phenotype discovery algorithm [37], originally developed by Yin et al. to filter image-based output for RNAi screens, was specifically adapted to identify putative organelle groupings. The resulting *phenoDisco* algorithm is shown schematically in Fig. 1. The following modelling protocol is performed independently over $N = 100$ iterations (determined empirically to yield stable solutions with practical CPU runtime) at the end of which new phenotypes may be identified. Upon identification of a new phenotype the phenotype class is added to the training data and the algorithm is restarted. The modelling protocol is once again repeated over $N = 100$ iterations. This protocol is repeated until no new phenotypes are identified.

The algorithm starts by selecting randomly without replacement a known phenotype dataset, g_m and combining it with all unlabelled data D_U, to obtain a combined subset $F = g_m \cup D_U$. The dataset F is then modelled using a Gaussian Mixture Model (GMM) (see Supplementary methods) using the Expectation Maximisation (EM) algorithm (as described in [46]) to identify proteins whose localisation profiles inhabit the same mixture components as those known to belong to the current class, m (stage 1, Fig. 1). These proteins are considered candidates for merging with the current phenotype class. The cluster number in the GMM is estimated using the Bayesian Information Criterion (BIC).

In Yin's original method, candidates identified are subsequently validated using a statistical test with Bonferroni correction to determine whether they will become new members of the current class. However, in the *phenoDisco* algorithm candidates from D_U are considered against the existing phenotype class via a non-parametric outlier detection test against a multivariate mixture [47] (described in Supplementary methods) (stage 2, Fig. 1). Candidates deemed sufficiently similar (rejection at 5%) are removed from D_U and merged with the current class in this iteration and those rejected remain unlabelled and are returned to D_U (stage 3, Fig. 1). The procedure is repeated for all organelle classes, $m = \{1,\ldots,K_0\}$, at the end of which concludes one full iteration.

In order to identify new phenotypes within the unlabelled data D_U, examples in D_U that are consistently accepted into a single class m over the 100 iterations are labelled as belonging to that class. Of the remaining unlabelled examples, those clustered together throughout each of the individual iterations but not incorporated into any known organelle phenotypes, are considered members of 'new' phenotypes (stage 4, Fig. 1). New phenotypes are added as new dataset classes to the pool of labelled data (augmenting the training data by incorporation of phenotypic examples). Upon the occurrence and definition of new phenotype classes the algorithm is restarted and repeated until no new phenotypes are defined.

3. Results and discussion

3.1. Novelty detection validation based on re-discovery of existing phenotypes

Validation of the phenotype discovery algorithm was performed by removing known organellar phenotypes from a dataset and assessing its ability to 're-discover' them. Given a dataset containing m known organelle clusters, m experiments were conducted in which one of the m clusters was deleted systematically from the training data prior to phenotype discovery. The algorithm displayed mean sensitivities of 0.926, 0.941 and 0.671 for the *Arabidopsis*, *Drosophila* and Human and datasets respectively, in correctly identifying examples from each missing phenotype as 'novel' (true positive), where

$$\text{Sensitivity} = \frac{\text{number of True Positives}}{\text{number of True Positives} + \text{number of False Negatives}}.$$

The lower sensitivity value for the Human dataset results from the lack of separation between the Golgi apparatus and endoplasmic reticulum (ER) compartments during density centrifugation. The algorithm struggles to re-find the Golgi in this example which lowers the mean sensitivity score. The specificity could not be computed in this set of experiments as the true number of false positives cannot be calculated. Any proteins identified as markers but which were not labelled prior

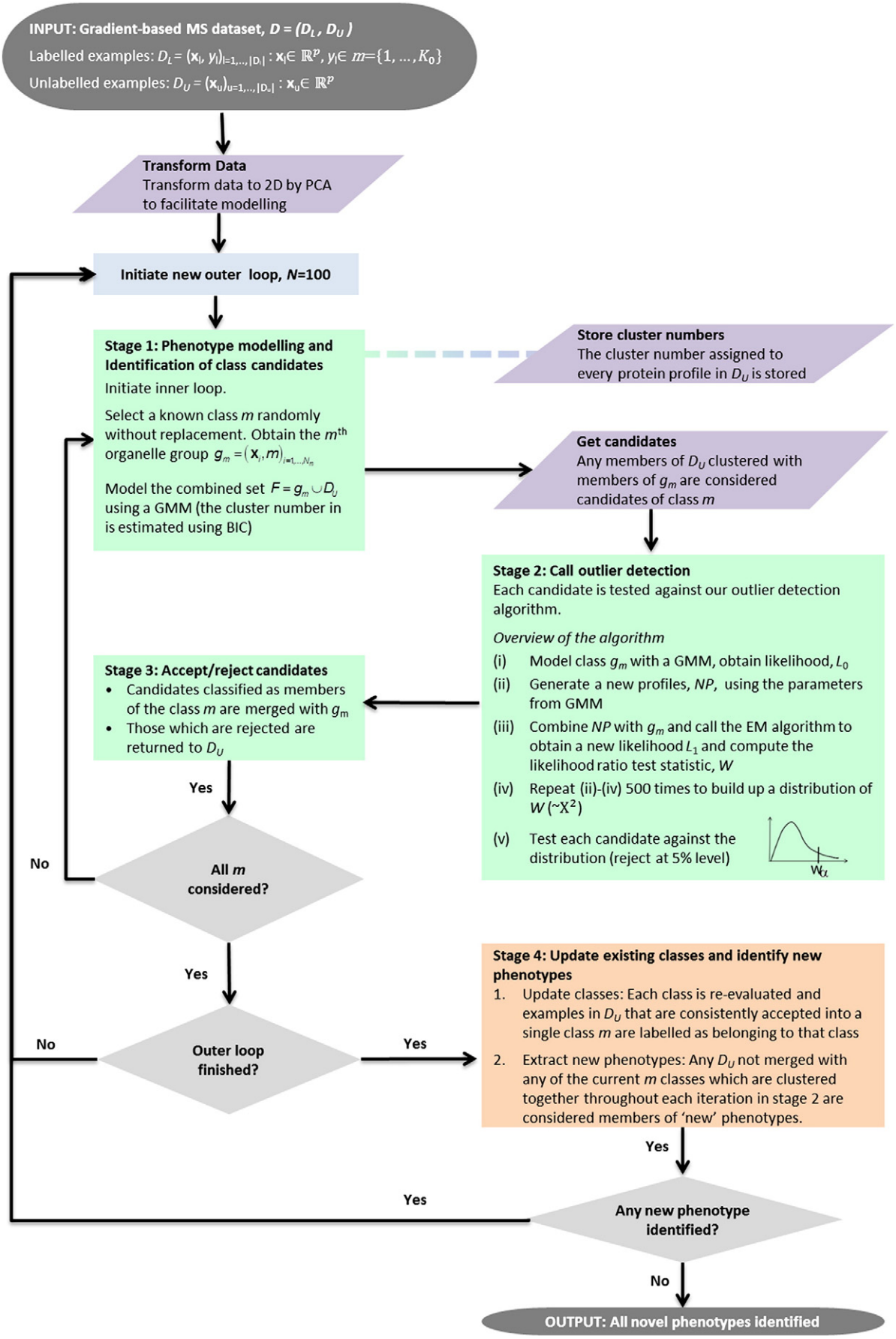

Fig. 1 – Simple workflow of the phenotype discovery algorithm.

to organelle discovery, i.e. were previously unknown, cannot be labelled as false positives as it is possible that they are in fact new members of the class under investigation.

This assessment also reveals an ability to identify organelle sub-compartments during their re-discovery. At the time of original publication of the *Arabidopsis* callus dataset [13], it was not possible to distinguish between the mitochondria and plastids. Accordingly, they appear as one class in the labelled training data (Fig. 2A). When this class is removed from the dataset, the discovery algorithm identifies a plastid cluster which is distinct from the mitochondria (Fig. 2B). The profiles of these plastid and mitochondria clusters exhibit distinct gradient distributions from one another (Supporting Fig. 1).

Similarly, upon removing the ER cluster, the discovery algorithm identified two separate ER phenotypes (Fig. 2A). One cluster consists of purely ER residents, whilst the second smaller cluster consists of largely of soluble proteins of which 9 of the 11 proteins in this cluster are PDI or PDI-like that are known to be localised to the ER lumen and contain the highly conserved tetrapeptide H/KDEL C-terminal retention sequence which constitutes the ER retention signal [48,49]. The tetrapeptide is recognised by the ERD2 receptor on the Golgi complex, resulting in the retrieval of H/KDEL proteins from this compartment back to the ER. Again, the normalised gradient distribution profiles of these two clusters are distinct from one another and represent the steady state distribution of this class of proteins which shuttle between ER and Golgi (Supporting Fig. 2). The other 2 proteins in this cluster have unknown annotation.

Finally, in Fig. 2A, phenotype 8 is identified by the *phenoDisco* algorithm as a specific group, distinct from the composite mitochondria/plastids marker set. If this 'artificial' joint mitochondria/plastid group is removed and the dataset resubmitted to phenotype discovery analysis, distinct groups corresponding to mitochondria and plastids are determined (Fig. 2B), and previously observed phenotype 8 is merged with the chloroplastic group. Similarly, we make the same observation with phenotype 5 in Fig. 2B when the ER cluster is removed (Fig. 2A). These cases illustrate differences in cluster definition under supervised (when the marker group involved is present and used to define an existing phenotype) and unsupervised (when it is not present and subsequently reconstituted) conditions respectively. Poorly defined organellar phenotypes, such as the combined mitochondria/plastid group, challenge the biological relevance of association in these circumstances and serve to highlight the importance of good initial marker annotation, whether applying the *phenoDisco* algorithm or any machine learning approach.

Further to the observations above, the deletion and subsequent rediscovery of known organelle markers also provide a means by which to test the suitability of applying the discovery approach in wider circumstances, especially to scenarios wherein class membership of starting organelles is either low in

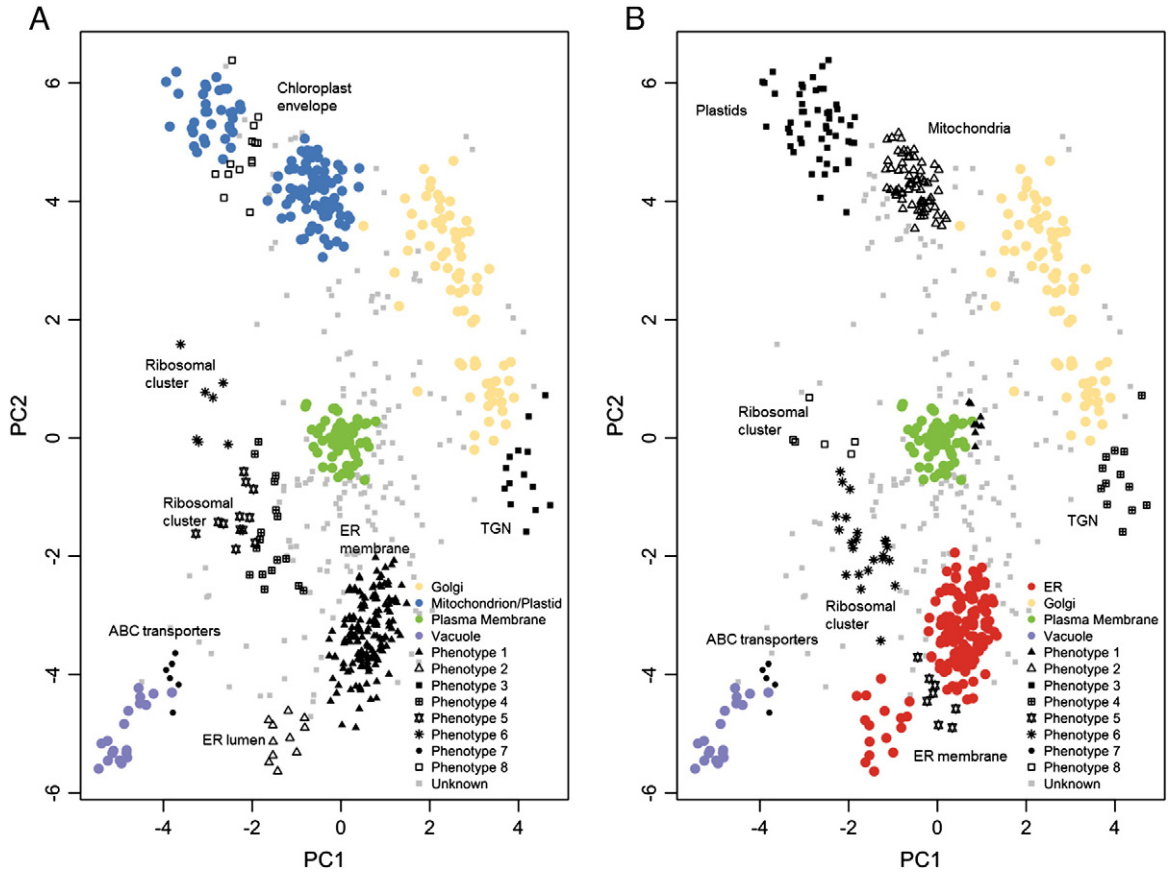

Fig. 2 – Principal component analysis (PCA) plots of LOPIT profiles of the *Arabidopsis thaliana* dataset [13] showing the clustering of proteins according to their density gradient distributions. (A) The ER was removed in this experiment and re-found as two separate clusters (phenotypes 1 and 2). (B) The combined mitochondrial and plastid cluster was removed in this experiment and re-found as two separate clusters (phenotypes 2 and 3).

number and/or strongly imbalanced. It is difficult to define a 'minimum use' scenario, or determine the tolerance of large imbalance in initial cluster membership, in a principled manner a priori because such properties tend to be data dependent. Rather than attempting to derive hard rules for appropriate use, however, the recovery of known organelle phenotypes provides an arguably more convincing and practical demonstration of applicability prior to discovery of new organelle phenotypes, and the practice is recommended in any future application to a wider range of datasets.

3.2. *Identification of putative organelle clusters and protein complexes*

Having observed that the discovery algorithm not only correctly identifies known organelle groupings removed from the data, but is also able to reconstitute them at greater resolution, the algorithm was applied to identify organelle phenotypes in three different LOPIT datasets [13,38]. All phenotype clusters discovered were examined and labelled manually by consulting the UniProt knowledgebase which collates a broad selection of curated and electronic annotation relating to protein function [50,51]. Manual querying of the database allowed thorough examination of protein type and protein function of annotated localisation (if given) to examine new phenotypes for biological relevance. Highlights of this application are described below, with full lists of phenotypes inferred from the unlabelled data of each dataset included in Supplementary information.

3.2.1. *The trans-Golgi network*

In the *Arabidopsis* callus dataset [13], 8 additional phenotype clusters were identified in the unlabelled portion of the data (Fig. 3A). Amongst these clusters was the putative trans-Golgi network (TGN) organelle (Fig. 3A; phenotype 2), which in plants functions as a hub for secretory and endocytic trafficking routes and acts to direct proteins to different subcellular locations [52]. The profiles for proteins belonging to this phenotype cluster exhibit a distinct density gradient distribution from proteins residing in the Golgi apparatus (Supporting Fig. 3).

The canonical view of the TGN is that it sits at the convergence of many trafficking routes, and the discrimination of cargo proteins in transit from functional residents of this organelle is challenging using traditional approaches. As such, there is a lack of robust protein markers described in experimental literature for the purpose of protein–organelle annotation. The TGN cluster identified here by *phenoDisco* contains the proteins ECHIDNA (At1g09330) and Syntaxin-43 (At3g05710), both widely-known TGN localised proteins [53,54]. Interestingly, upon validation of the remaining proteins it is found that 8 have been associated with the TGN (e.g. At4g12650, At4g30260, At5g64030) in recent Syp61 vesicle immunoisolation experiments by Drakakaki et al. [55] and the remaining have been associated as TGN or GA localised in experiments by Nikolovski et al. but their location to one or the other location not distinguished to date [56] (see Supporting information). Members of the vacuolar-sorting receptor families are present amongst these proteins and it has recently been proposed that the TGN is the location of retromer-mediated recycling of VSRs [57]. It is also worthy of note that, with the exception of ECHIDNA and Syntaxin-43, these proteins are not widely known as markers or at all well-annotated in the literature, which highlights the application of *phenoDisco* in discovering potential new markers for further evaluation.

3.2.2. *Other organelles and sub-compartments*

A well-defined lysosomal protein cluster was also identified in the Human dataset (Fig. 3C; phenotype 3) wherein 10 of the 17 of proteins in this cluster are well-known markers of this subcellular niche. Many of the proteins classified to this phenotypic group are integral to the function of this organelle, including several proteolytic hydrolases such as the acid protease Cathepsin D [58], and the cation-independent mannose-6-phosphate receptor, which has a well-established role in the trafficking of nascent hydrolases to the lysosomal compartment [59]. Of the remaining 7 proteins in this phenotype it was found that 5 proteins had unknown annotation and 2 proteins had an ambiguous undetermined localisation.

Other organelles, such as peroxisomes and lysosomes, were discovered in both the *Drosophila* and Human datasets. Well-known peroxisomal marker catalase [60], along with acyl-coenzyme A oxidase and peroxisomal multifunctional enzymes type 2 (Mfe2) was detected in the *Drosophila* dataset. In the Human dataset a small cluster of 5 proteins was detected (phenotype 7) which also contains proteins with peroxisomal annotation, including 3-ketoacyl-CoA thiolase [60] and alkylglycerone-phosphate synthase [61].

3.2.3. *Protein complexes*

Although the LOPIT experiments described here were designed to separate organelles of interest and classify their respective protein residents, the *phenoDisco* algorithm also identifies several phenotypes relating to non-organellar macromolecular protein complexes, such as ribosomes and proteasomes, within all three datasets. This suggests that these complexes are sufficiently dense to exhibit distinct density gradient distributions that enable their detection and resolution from other subcellular niches. Small clusters containing members of the small (40S) and large (60S) ribosomal subunits are found in both the *Arabidopsis* and *Drosophila* datasets (Fig. 3A; phenotypes 4 and 6, Fig. 3B; phenotypes 1 and 3). In the case of the *Drosophila* dataset, 25 proteins in phenotype 1 are known members of the large ribosomal subunit, and 14 proteins in phenotype 3 are recognised small ribosomal proteins. Interestingly, the heavily enriched large ribosomal cluster in the *Drosophila* dataset (phenotype 1) also contains 8 proteins that are thought to be lysosomal residents. This could be due to a lack of separation of these two subcellular niches by the fractionation scheme used in this experiment, with the four density gradient fractions selected for analysis insufficient to resolve these compartments. Although the identified phenotype is comprised of two different compartments due to limitations of the separation technique employed, the algorithm successfully highlights organelles that may have been otherwise overlooked or misclassified as a result of insufficient annotation. Feedback from this type of analysis thus facilitates optimisation of subsequent organelle separation experiments. The full list of proteins, including protein names and known localisations can be found in the Supporting information.

Consistent with Tan et al. a sub-cluster of proteasome localised proteins were found in phenotype 2 of the *Drosophila* data. Proteasomes are multi-subunit protein complexes located

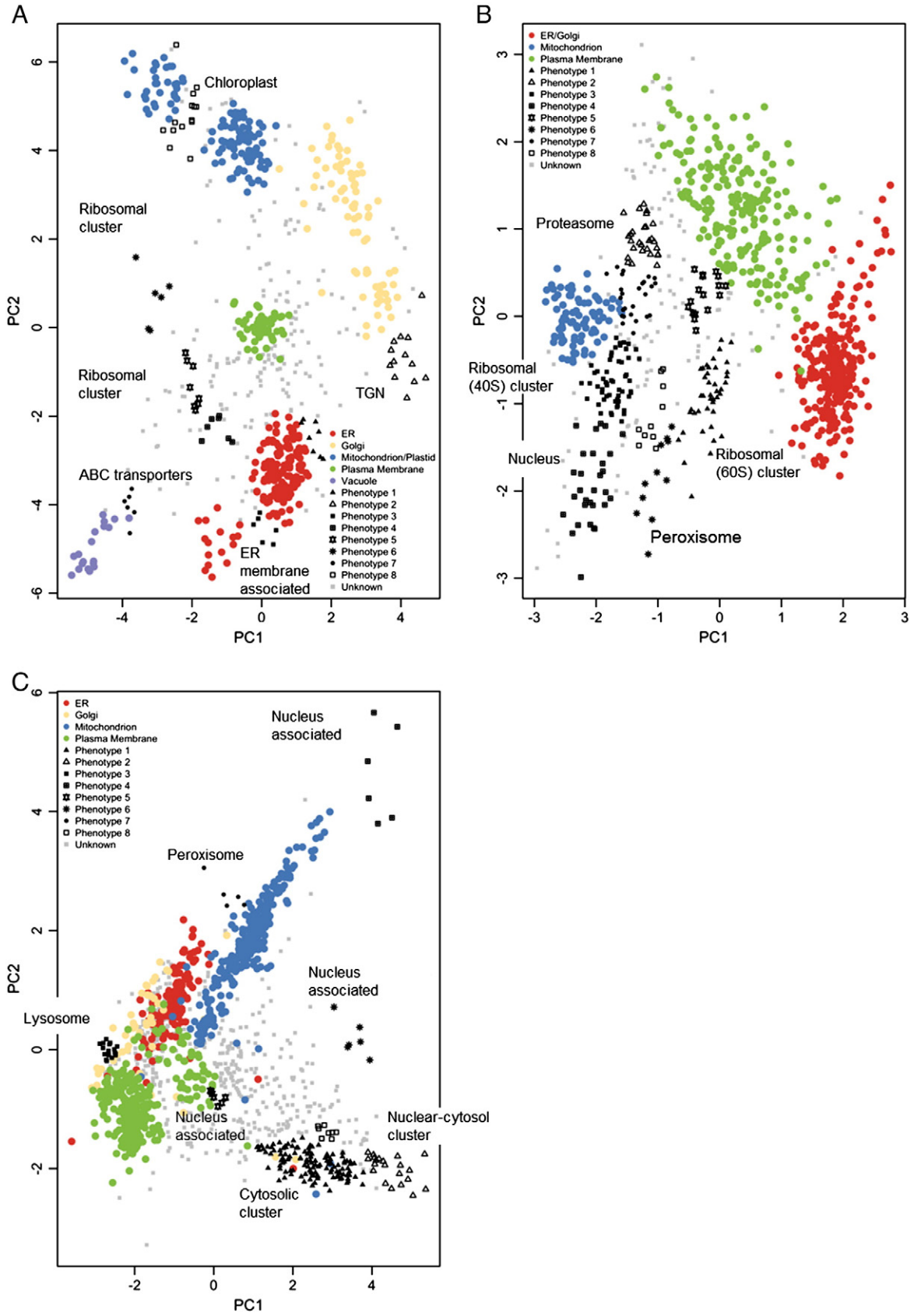

Fig. 3 – Principal component analysis (PCA) plots of LOPIT profiles showing the clustering of proteins according to their density gradient distributions in (A) an *Arabidopsis thaliana* dataset [13], (B) a *Drosophila melanogaster* dataset [28], and (C) of a Human HEK 293 dataset. Putative organelles and protein complexes assigned from the phenotype discovery algorithm are highlighted.

in the nucleus and cytosol which are the effectors for one of the primary routes of protein degradation within the cell. It is found that the majority of proteins identified in phenotype 2 belong to the 26S proteasome subunit. The *phenoDisco* algorithm also identifies a number of cytosolic complexes in the *Drosophila* dataset including a cluster of cytoskeletal associated proteins (Fig. 3B; phenotype 7) of which the majority specifically microtubule associated.

3.2.4. *Identification of experimental artefacts*

The identification of nucleus and associated sub-compartments of the nucleus in the *Drosophila* and Human datasets highlights the potential utility of identifying compartments that the experiment was not designed to enrich for. In both experiments, the nuclei were discarded by centrifugation at an early stage of the experimental schema, though residual quantities of this organelle persisted due to the difficulties associated with achieving total purification or extraction of an organelle, and the high abundance of some nuclear proteins. This demonstrates a distinct advantage of combining multivariate analytical fractionation strategies such as LOPIT with phenotype discovery analysis. Rather than these nuclear proteins contaminating other sub-cellular fractions and being misclassified, they form their own characteristic fractionation distribution, which is then both detected and resolved from other phenotypes in the analysis despite there being no a priori knowledge of such an organelle and distribution being present in the sample.

Distinct clusters of nuclear-associated contaminant proteins were revealed in the *Drosophila* (Fig. 3B; phenotype 4) and Human datasets (Fig. 3C; phenotypes 2, 4, 5 and 6). The phenotype discovery analysis was able to identify a large cluster of proteins with mixed nuclear-cytosol localisation (Fig. 3C; phenotype 2), including nuclear export receptor exportin 2, mapmodulin, and the human orthologue of serrate, all of which have been previously observed to shuttle between the nuclear and cytosolic compartments [62–64].

3.2.5. *Protein families*

Interestingly, some of the phenotype clusters identified correspond to small clusters of proteins from the same family. For example, in the *Arabidopsis* dataset we find phenotype 7 (Fig. 3A) contains several members and putative members of the ATP-binding cassette (ABC) transporter families B and C. The identification of these proteins gives additional information about their sub-cellular localisation. ABC transporters are one of the largest families of transmembrane proteins. These proteins transport of various molecules across the cell membrane in an ATP dependent manner. They are known to be localised in many different organelles such as plasma membrane, mitochondria, chloroplasts, vacuolar membrane and peroxisomes. Although the plant genome encodes more than 130 ABC transporters, in *Arabidopsis* only 22 have been functionally analysed and thus there is still little known about many ABC transporter sub-families [65].

The ABC transporters ABCB27 (TAP2), ABCC3 and ABCC8 are known to be localised to the vacuolar membrane [66–68] and are found in a distinct cluster (Fig. 3A, phenotype 7). Three other proteins were also found in this phenotype cluster; a major facilitator family protein (At2g1660), a non-specific phospholipase C protein (At3g03520) and a putative uncharacterised protein (At4g38350). The fact that phenotype 7 is found distinct from the vacuole is interesting and implies that their steady state is not entirely vacuolar which is suggestive of involvement of these proteins in concerted recycling. Other ABC transporter proteins identified in the *Arabidopsis* dataset largely localise to the plasma membrane or vacuole (Table 2). Interestingly, some ABC transporters (ACBB6, ABCG36 and 14 and ABCC14) which are annotated as plasma membrane proteins do not co-cluster with other proteins residing in this organelle. This may reflect their dynamic nature and the fact that their steady state localisation is influenced by their residence at several sub-cellular locations.

Another small cluster in the *Arabidopsis* dataset; phenotype 3, which contains predominantly ER localised proteins, lies between the two large clusters that comprise the ER. As discussed in Section 3.1 if the ER markers are removed from the starting labelled training data and then the dataset is resubmitted to the *phenoDisco* analysis, two very distinct ER lumen and ER membrane-associated protein clusters are uncovered and we find phenotype 3 (Fig. 3A), which is enriched with cytochrome P450 oxidases, is then merged with the ER membrane cluster. This highlights not only the impact that the level of annotation can have on cluster analysis, but also that the *phenoDisco* algorithm has the capability to identify phenotypes within the data with a level of intra-organellar resolution that surpasses conventional annotation of protein subcellular localisation.

The biological relevance of protein groups identified via phenotype discovery demonstrate that not all of the proteins in each dataset correspond to organelles represented by the training data used in previous applications of supervised classification approaches. In this circumstance, it is obvious that proteins in these newly-identified groups would be misclassified as a result of a classifier trained on fewer organelle phenotypes failing to incorporate a decision boundary that distinguishes them. These observations suggest, therefore, the importance of identifying and labelling as many organelle phenotypes as possible before seeking to associate any proteins without organelle annotation. It is also clear that many of the groups identified by phenotype discovery comprise small numbers of protein profiles that may not be suitable for inclusion by supervised classification. In turn, this suggests the inclusion of a single, more generic data class of small phenotypes. Proteins associated to this class within a supervised classification framework would provide candidates for further investigation (similar to the above) and would at least avoid erroneous classification into one of the larger organelle phenotypes obtained during initial data annotation.

4. Conclusions

In this study, we have addressed a key issue that hinders current protein localisation prediction approaches in gradient-based MS data. Current methods largely use supervised machine learning algorithms to map profiles of relative protein abundance along a gradient to known protein–organelle associations. Such approaches are highly likely to make association errors if the training data does not include all sub-cellular compartments present in the experimental data. Mining of such datasets is laborious and can be difficult due to the lack of known markers in

Table 2 – ABC transporter proteins detected in the *Arabidopsis* callus LOPIT dataset [13].

TAIR locus [41]	Protein name	*PhenoDisco* localisation	Localisation [65]
At5g58270	ABC transporter B family member 25, mitochondrial	Mitochondria	Mitochondria
At3g13080	ABC transporter C family member 3 (ABCC3)	Phenotype 7	Vacuole
At3g21250	ABC transporter C family member 8 (ABCC8)	Phenotype 7	Vacuole
At5g39040	ABC transporter B family member 27 (TAP2)	Phenotype 7	Vacuole
At1g02520	ABC transporter B family member 4 (ABCB4)	Plasma membrane	Plasma membrane
At1g15210	ABC transporter G family member 35 (ABCG35)	Plasma membrane	Plasma membrane
At2g36380	ABC transporter G family member 34 (ABCG34)	Plasma membrane	Plasma membrane
At2g36910	ABC transporter B family member 1 (ABCB1)	Plasma membrane	Plasma membrane
At2g47000	ABC transporter B family member 4 (ABCB4)	Plasma membrane	Plasma membrane
At3g53480	ABC transporter G family member 37 (ABCG37)	Plasma membrane	Plasma membrane
At1g59870	ABC transporter G family member 36 (ABCG36)	Unknown	Plasma membrane
At2g39480	ABC transporter B family member 6 (ABCB6)	Unknown	Plasma membrane
At2g47800	ABC transporter C family member 4 (ABCC4)	Unknown	Vacuole
At3g28860	ABC transporter B family member 19 (ABCB19)	Unknown	Plasma membrane
At3g62700	ABC transporter C family member 14 (ABCC14)	Unknown	Vacuole
At1g04120	ABC transporter C family member 5 (ABCC5)	Vacuole	Vacuole
At1g30400	ABC transporter C family member 1 (ABCC1)	Vacuole	Vacuole
At2g34660	ABC transporter C family member 2 (ABCC2)	Vacuole	Vacuole

databases and the literature. Here, we have presented an adaptation of a published phenotype discovery algorithm [37] which is used to identify putative organelle clusters in gradient-based MS datasets and which appears highly effective in the scenario presented.

Three different empirical datasets produced from LOPIT [13,16] experiments were adopted in our investigation. A number of organelles, sub-compartments and protein complexes were identified, many of which were not known to be present in the data prior to analysis and as such were not represented in the original training datasets used previously for protein localisation prediction experiments. A number of additional well-known organelles, such as lysosomes and peroxisomes, were also identified during this new analysis, alongside (1) putative organelle sub-compartments, such as the trans-Golgi network, (2) protein complexes, such as ribosomes and proteasomes, and (3) small clusters of protein families, such as ABC transporters. Furthermore, contaminant clusters that the experiments were not designed to enrich for, such as nucleus and associated sub-compartments, were identified in both *Drosophila* and Human datasets, thereby highlighting the distinct advantages of incorporating novelty detection analysis in gradient-based MS strategies.

The phenotype discovery approach presented was carefully validated through assessment of its ability to re-discover known phenotype clusters when systematically deleted from the training data. The protein content of new phenotype clusters was examined extensively and localisation information was retrieved from online databases and the literature to assess biological relevance. It is found that by employing an organelle discovery approach, one is able to mine gradient-based MS datasets at a deeper level and bring to light interesting sub-cellular compartments and protein complexes for more comprehensive validation.

When considering what the results obtained mean for present analytical practice on this type of organelle proteomics data, it is clear that failure to extract all organelle classes present in the data leads supervised classification methods to train on a limited set of organelle phenotypes beyond which unannotated proteins may not be associated. This will obviously lead to erroneous organelle associations. Accordingly, the results suggest a substantial benefit to employing an organelle discovery approach prior to the application of standard supervised classification approaches. Moreover, this may be expanded to suggest the inclusion of data structure prior to *or during* the creation of protein–organelle classifiers on sub-cellular fractionation based data, and the explicit application of *semi-supervised* machine learning approaches in order to do so.

In this regard, although protein localisation prediction from these data has to date been limited to supervised learning methods, the prediction of protein localisation from amino acid sequence is a mature field and a wide variety of classification approaches have been employed that make use of supervised, unsupervised and semi-supervised learning. A comprehensive recent review [7] outlines the extensive application of machine learning to sequence data to predict protein–organelle associations, and semi-supervised learning has been applied to organelle discovery from sequence-based representation [69,70]. With the growing amount of data produced from gradient-based experiments coupled with the limited labelled training data available in the literature and associated databases, the results presented here suggest that similar semi-supervised approaches represent an interesting direction for future work in the computational association of proteins to organelles from organelle proteomics data.

In summary, phenotype discovery analysis applied to sub-cellular localisation data has revealed a number of interesting new phenotype clusters which represent organelles, organellar sub-compartments, protein complexes and protein families within the datasets examined. The algorithm applied was able to identify both large and small organelle-related protein clusters beyond those sufficiently annotated for use as training data in present supervised classification approaches to protein–organelle association. Some of the small phenotype clusters, many of which are too small to be used as training data for generalisable machine learning solutions, present candidates for more comprehensive validation. Rather than present an alternative to supervised

classification approaches, however, the method presented enables deeper mining of such datasets post-annotation but prior to predictive organelle association, and extracts interesting phenotype clusters for subsequent analysis or more comprehensive evaluation in an unbiased fashion. The phenotype discovery algorithm is implemented in the R programming language and is readily available, along with state-of-the-art supervised machine learning procedures and example datasets, in the Bioconductor [45] *pRoloc* (http://bioconductor.org/packages/devel/bioc/html/pRoloc.html) and *pRolocdata* (http://bioconductor.org/packages/devel/data/experiment/html/pRolocdata.html) packages. Finally, the results obtained suggest the future application of semi-supervised machine learning approaches to similar gradient-based localisation datasets, in order to incorporate similarities identified amongst un-annotated protein profiles prior to or during construction of organelle-based classifiers.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jprot.2013.02.019.

Acknowledgements

The authors would like to thank Dr. Sean Holden, University of Cambridge Computer Laboratory, for helpful discussions. This work was primarily funded by BBSRC grant BB/H024247/1 which supported LMB as a postdoctoral researcher. LG is supported by a 7th Framework Programme of the European Union (262067-PRIME-XS). AC was supported by BBSRC grant BB/D526088/1. AJG was funded by generous funding from the King Abdullah University for Science and Technology, Saudi Arabia. MWBT is an employee of Celgene Research SLU (Spain), part of Celgene Corporation.

REFERENCES

[1] Gatto L, Vizcaíno JA, Hermjakob H, Huber W, Lilley KS. Organelle proteomics experimental designs and analysis. Proteomics 2010;10:3957–69.
[2] Dreger M. Subcellular proteomics. Mass Spectrom Rev 2003;22:27–56.
[3] Lilley KS, Dupree P. Plant organelle proteomics. Curr Opin Plant Biol 2007;10:594–9.
[4] Chou K-C, Shen H-B. Recent progress in protein subcellular location prediction. Anal Biochem 2007;370:1–16.
[5] Casadio R, Martelli PL, Pierleoni A. The prediction of protein subcellular localization from sequence: a shortcut to functional genome annotation. Brief Funct Genomic Proteomic 2008;7:63–73.
[6] Imai K, Nakai K. Prediction of subcellular locations of proteins: where to proceed? Proteomics 2010;10:3970–83.
[7] Du P, Li T, Wang X. Recent progress in predicting protein sub-subcellular locations. Expert Rev Proteomics 2011;8:391–404.
[8] Huh W-K, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, et al. Global analysis of protein localization in budding yeast. Nature 2003;425:686–91.
[9] Fagerberg L, Stadler C, Skogs M, Hjelmare M, Jonasson K, Wiking M, et al. Mapping the subcellular protein distribution in three human cell lines. J Proteome Res 2011;10:3766–77.
[10] Gilchrist A, Au CE, Hiding J, Bell AW, Fernandez-Rodriguez J, Lesimple S, et al. Quantitative proteomics analysis of the secretory pathway. Cell 2006;127:1265–81.
[11] Lam YW, Lamond AI, Mann M, Andersen JS. Analysis of nucleolar protein dynamics reveals the nuclear degradation of ribosomal proteins. Curr Biol 2007;17:749–60.
[12] Andreyev AY, Shen Z, Guan Z, Ryan A, Fahy E, Subramaniam S, et al. Application of proteomic marker ensembles to subcellular organelle identification. Mol Cell Proteomics 2010;9:388–402.
[13] Dunkley TPJ, Hester S, Shadforth IP, Runions J, Weimar T, Hanton SL, et al. Mapping the *Arabidopsis* organelle proteome. PNAS 2006;103:6518–23.
[14] Foster LJ, De Hoog CL, Zhang Y, Zhang Y, Xie X, Mootha VK, et al. A mammalian organelle map by protein correlation profiling. Cell 2006;125:187–99.
[15] Wiese S, Gronemeyer T, Ofman R, Kunze M, Grou CP, Almeida JA, et al. Proteomics characterization of mouse kidney peroxisomes by tandem mass spectrometry and protein correlation profiling. Mol Cell Proteomics 2007;6:2045–57.
[16] Sadowski PG, Dunkley TP, Shadforth IP, Dupree P, Bessant C, Griffin JL, et al. Quantitative proteomic approach to study subcellular localization of membrane proteins. Nat Protoc 2006;1:1778–89.
[17] Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, et al. Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. Mol Cell Proteomics 2004;3:1154–69.
[18] Thompson A, Schäfer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, et al. Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Anal Chem 2003;75:1895–904.
[19] Harner M, Körner C, Walther D, Mokranjac D, Kaesmacher J, Welsch U, et al. The mitochondrial contact site complex, a determinant of mitochondrial architecture. EMBO J 2011;30:4356–70.
[20] Dengjel J, Jakobsen L, Andersen JS. Organelle proteomics by label-free and SILAC-based protein correlation profiling. Methods Mol Biol 2010;658:255–65.
[21] De Duve C. Tissue fractionation. Past and present. J Cell Biol 1971;50:20d–55d.
[22] Andersen JS, Wilkinson CJ, Mayor T, Mortensen P, Nigg EA, Mann M. Proteomic characterization of the human centrosome by protein correlation profiling. Nature 2003;426:570–4.
[23] Vapnik VN. The nature of statistical learning theory. Springer; 2000.
[24] Cristianini N, Shawe-Taylor J. An introduction to support vector machines: and other kernel-based learning methods. Cambridge University Press; 2000.
[25] Trotter MWB, Sadowski PG, Dunkley TPJ, Groen AJ, Lilley KS. Improved sub-cellular resolution via simultaneous analysis of organelle proteomics data across varied experimental conditions. Proteomics 2010;10:4213–9.
[26] Zhang Y, Li T, Yang C, Li D, Cui Y. Prelocabc: a novel predictor of protein sub-cellular localization using a Bayesian classifier. J Proteomics Bioinform 2011;04.
[27] Chou K-C, Shen H-B. Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. J Proteome Res 2007;6:1728–34.
[28] Shen H-B, Chou K-C. Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. Biochem Biophys Res Commun 2007;355:1006–11.
[29] Mitschke J, Fuss J, Blum T, Höglund A, Reski R, Kohlbacher O, et al. Prediction of dual protein targeting to plant organelles. New Phytol 2009;183:224–35.
[30] Yang Y, Lu B-L. Protein subcellular multi-localization prediction using a min-max modular support vector machine. Int J Neural Syst 2010;20:13–28.

[31] Xiao X, Wu Z-C, Chou K-C. A multi-label classifier for predicting the subcellular localization of gram-negative bacterial proteins with both single and multiple sites. PLoS One 2011;6:e20592.
[32] Wan S, Mak M-W, Kung S-Y. mGOASVM: multi-label protein subcellular localization based on gene ontology and support vector machines. BMC Bioinformatics 2012;13:290.
[33] Briesemeister S, Rahnenführer J, Kohlbacher O. YLoc—an interpretable web server for predicting subcellular localization. Nucleic Acids Res 2010;38:W497–502.
[34] Briesemeister S, Rahnenführer J, Kohlbacher O. Going from where to why—interpretable prediction of protein subcellular localization. Bioinformatics 2010;26:1232–8.
[35] Li L-Q, Zhang Y, Zou L-Y, Zhou Y, Zheng X-Q. Prediction of protein subcellular multi-localization based on the general form of Chou's pseudo amino acid composition. Protein Pept Lett 2012;19:375–87.
[36] Mei S. Multi-label multi-kernel transfer learning for human protein subcellular localization. PLoS One 2012;7:e37716.
[37] Yin Z, Zhou X, Bakal C, Li F, Sun Y, Perrimon N, et al. Using iterative cluster merging with improved gap statistics to perform online phenotype discovery in the context of high-throughput RNAi screens. BMC Bioinformatics 2008;9:264.
[38] Tan DJL, Dvinge H, Christoforou A, Bertone P, Martinez Arias A, Lilley KS. Mapping organelle proteins and protein complexes in *Drosophila melanogaster*. J Proteome Res 2009;8:2667–78.
[39] The UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res 2011;40:D71–5.
[40] McQuilton P, St Pierre SE, Thurmond J. FlyBase 101—the basics of navigating FlyBase. Nucleic Acids Res 2012;40:D706–14.
[41] Garcia-Hernandez M, Berardini TZ, Chen G, Crist D, Doyle A, Huala E, et al. TAIR: a resource for integrated *Arabidopsis* data. Funct Integr Genomics 2002;2:239–53.
[42] Swarbreck D, Wilks C, Lamesch P, Berardini TZ, et al. The *Arabidopsis* Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res 2007;36:D1009–14.
[43] Mi H, Dong Q, Muruganujan A, Gaudet P, Lewis S, Thomas PD. PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. Nucleic Acids Res 2010;38:D204–10.
[44] Gatto L, Lilley KS. MSnbase—an R/bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. Bioinformatics 2012;28:288–9.
[45] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004;5:R80.
[46] Zhou X, Wang X. Optimisation of Gaussian mixture model for satellite image classification. IEE Proc Vis Image Signal Process 2006;153:349–56.
[47] Wang S, Woodward WA, Gray HL, Wiechecki S, Sain SR. A new test for outlier detection from a multivariate mixture distribution. J Comput Graph Stat 1997;6:285–99.
[48] Munro S, Pelham HR. A C-terminal signal prevents secretion of luminal ER proteins. Cell 1987;48:899–907.
[49] Pelham HR. Control of protein exit from the endoplasmic reticulum. Annu Rev Cell Biol 1989;5:1–23.
[50] Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. UniProt: the universal protein knowledgebase. Nucleic Acids Res 2004;32:D115–9.
[51] Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. The universal protein resource (UniProt). Nucleic Acids Res 2005;33:D154–9.
[52] Griffiths G, Simons K. The trans Golgi network: sorting at the exit site of the Golgi complex. Science 1986;234:438–43.
[53] Gendre D, Oh J, Boutté Y, Best JG, Samuels L, Nilsson R, et al. Conserved *Arabidopsis* ECHIDNA protein mediates trans-Golgi-network trafficking and cell elongation. PNAS 2011;108:8048–53.
[54] Uemura T, Ueda T, Ohniwa RL, Nakano A, Takeyasu K, Sato MH. Systematic analysis of SNARE molecules in *Arabidopsis*: dissection of the post-Golgi network in plant cells. Cell Struct Funct 2004;29:49–65.
[55] Drakakaki G, Van de Ven W, Pan S, Miao Y, Wang J, Keinath NF, et al. Isolation and proteomic analysis of the SYP61 compartment reveal its role in exocytic trafficking in *Arabidopsis*. Cell Res 2012;22:413–24.
[56] Nikolovski N, Rubtsov D, Segura MP, Miles GP, Stevens TJ, Dunkley TPJ, et al. Putative glycosyltransferases and other plant Golgi apparatus proteins are revealed by LOPIT proteomics. Plant Physiol 2012;160:1037–51.
[57] Niemes S, Langhans M, Viotti C, Scheuring D, San Wan Yan M, Jiang L, et al. Retromer recycles vacuolar sorting receptors from the trans-Golgi network. Plant J 2010;61:107–21.
[58] Sagné C, Agulhon C, Ravassard P, Darmon M, Hamon M, El Mestikawy S, et al. Identification and characterization of a lysosomal transporter for small neutral amino acids. PNAS 2001;98:7206–11.
[59] Sahagian GG. The mannose 6-phosphate receptor: function, biosynthesis and translocation. Biol Cell 1984;51:207–14.
[60] Glover JR, Andrews DW, Subramani S, Rachubinski RA. Mutagenesis of the amino targeting signal of Saccharomyces cerevisiae 3-ketoacyl-CoA thiolase reveals conserved amino acids required for import into peroxisomes in vivo. J Biol Chem 1994;269:7558–63.
[61] De Vet ECJ, Van den Broek BT, Van den Bosch H. Nucleotide sequence of human alkyl-dihydroxyacetonephosphate synthase cDNA reveals the presence of a peroxisomal targeting signal 2. Biochim Biophys Acta 1997;1346:25–9.
[62] Kutay U, Bischoff FR, Kostka S, Kraft R, Görlich D. Export of importin alpha from the nucleus is mediated by a specific nuclear transport factor. Cell 1997;90:1061–71.
[63] Tsujio I, Zaidi T, Xu J, Kotula L, Grundke-Iqbal I, Iqbal K. Inhibitors of protein phosphatase-2A from human brain structures, immunocytological localization and activities towards dephosphorylation of the Alzheimer type hyperphosphorylated tau. FEBS Lett 2005;579:363–72.
[64] Gilmore-Hebert M, Ramabhadran R, Stern DF. Interactions of ErbB4 and Kap1 connect the growth factor and DNA damage response pathways. Mol Cancer Res 2010;8:1388–98.
[65] Kang J, Park J, Choi H, Burla B, Kretzschmar T, Lee Y, et al. Plant ABC transporters. Arabidopsis Book 2011;9.
[66] Jaquinod M, Villiers F, Kieffer-Jaquinod S, Hugouvieux V, Bruley C, Garin J, et al. A proteomics dissection of *Arabidopsis thaliana* vacuoles isolated from cell culture. Mol Cell Proteomics 2006;6:394–412.
[67] Rea PA. Plant ATP-binding cassette transporters. Annu Rev Plant Biol 2007;58:347–75.
[68] Nagy R, Grob H, Weder B, Green P, Klein M, Frelet-Barrand A, et al. The *Arabidopsis* ATP-binding cassette protein AtMRP5/AtABCC5 is a high affinity inositol hexakisphosphate transporter involved in guard cell signaling and phytate storage. J Biol Chem 2009;284:33614–22.
[69] Xu Q, Hu DH, Xue H, Yu W, Yang Q. Semi-supervised protein subcellular localization. BMC Bioinformatics 2009;10:S47.
[70] Caragea C, Caragea D, Silvescu A, Honavar V. Semi-supervised prediction of protein subcellular localization using abstraction augmented Markov models. BMC Bioinformatics 2010;11:S6.