A semi-supervised Bayesian approach for simultaneous protein sub-cellular localisation and organelle discovery

Oliver M. Crook[1,2,3], Lisa M. Breckels[1,2], Kathryn S. Lilley[2], Paul D.W. Kirk[3], and Laurent Gatto* [4]

[1]Computational Proteomics Unit, Department of Biochemistry, University of Cambridge, Cambridge, UK
[2]Cambridge Centre for Proteomics, Department of Biochemistry, University of Cambridge, Cambridge, UK
[3]MRC Biostatistics Unit, Cambridge Institute for Public Health, Cambridge, UK
[4]de Duve Institute, UCLouvain, Avenue Hippocrate 75, 1200 Brussels, Belgium

December 2, 2018

Abstract

1 Introduction

Aberrant protein sub-cellular localisation has been implicated in numerous diseases, including cancers (Kau *et al.*, 2004), obesity (Siljee *et al.*, 2018), and multiple others (Laurila and Vihinen, 2009). Characterising the sub-cellular localisation of proteins is therefore of critical importance in order to understand the pathobiological mechanisms and aetiology of many diseases. Proteins are compartmentalised into sub-cellular niches, including organelles, sub-cellular structures and protein complexes. These compartments ensure the biochemical conditions for proteins to function correctly are met, as well as being in the proximity of intended interaction partners (Gibson, 2009). High-throughput and high-accuracy mass-spectrometry (MS) based methods to map the global sub-cellular landscape now exist (Christoforou *et al.*, 2016; Mulvey *et al.*, 2017; Geladaki *et al.*, 2018). These method begin with gentle cell lysis, proceeded by sub-cellular fractionation and MS-based proteomics profiling. These spatial

*laurent.gatto@uclouvain.be

proteomics approaches rely on rigorous data analysis and interpretation (Gatto *et al.*, 2010, 2014a).

Current computational approaches in MS-based spatial proteomics rely on machine learning algorithms to make protein-organelle associations, see (Gatto *et al.*, 2014a) for an overview. If a dataset is insufficiently annotated such that sub-cellular niches that are present in the experimental data are not known to the classifier then this leads to the classifier making erroneous assignments, resulting in inflated FDR and uncertainty estimates. Therefore, novelty detection and organelle discovery is of vital importance in spatial proteomics.

Novelty detection can also prove useful in validating experimental design, either by demonstrating contaminates have been removed or increased resolution of organelle classes. Furthermore, organisms with little *a priori* knowledge of their proteome organisation can be challenging to annotate and novelty detection can provide putative evidence for sufficient resolution. Since resolution is an important factor in deciding how the proteome should be annotated, Gatto *et al.* (2018) proposed a quantitative measure of organelle resolution to guide users.

Breckels *et al.* (2013) presented a phenotype discovery algorithm called *phenoDisco* to detect novel sub-cellular niches and alleviate the issue of undiscovered phenotypes. The algorithm uses a iterative procedure and the BIC (Schwarz *et al.*, 1978) is employed to decide the number of newly detected phenotypes. Afterwards the dataset can be re-annotated and a classifier employed to assign proteins to organelles. Breckels *et al.* (2013) applied their method on several datasets and discovered new organelle classes in *Arabidopsis* (Dunkley *et al.*, 2006) and *Drosophila* (Tan *et al.*, 2009). The approach later successfully identified the trans-golgi network (TGN) in *Arabidopsis* roots (Groen *et al.*, 2014).

Recently, Crook *et al.* (2018) demonstrated the importance of uncertainty quantification in spatial proteomics. They proposed a generative classifier and took a fully Bayesian approach to spatial proteomics data analysis by computing probability distributions of protein-organelle assignments using Markov-chain Monte-Carlo (MCMC). These probabilities are then used as the basis of organelle allocations, as well as quantifying the uncertainty in these allocations. Some proteins are not well described by any of the annotated sub-cellular niches and so a multivariate student's T distribution is included for outlier detection. The proposed T-Augmented Gaussian Mixture (TAGM) model also achieves state-of-the-art predictive performance against other commonly used machine learning algorithms (Crook *et al.*, 2018).

One potential source of uncertainty is undetected or un-annotated sub-cellular niches. Thus, we propose an extension to TAGM to allow simultaneous protein-organelle assignments and novelty detection. One assumption of the TAGM model is that the number of organelle classes is known, and thus we design a novelty detection algorithm based on allowing an unknown number of organelle classes, as well as quantifying uncertainty in this number.

Quantifying uncertainty in the number of components in a Bayesian mixture model is challenging and most approaches either rely on a non-parametric prior called the Dirichlet process (Ferguson, 1974; Antoniak, 1974) or reversible-jump Markov-chain Monte-Carlo (RJ MCMC) approaches (Richardson and Green, 1997). Here we make use of recent asymptotic results in Bayesian analysis of mixture models (Rousseau and Mengersen, 2011). The principle of overfitted mixtures allows us to specify a (possibly large) maximum number of components. As shown in Rousseau and Mengersen (2011) these components empty if they

are not supported by the data, allowing the number of components to be inferred.

Bayesian approaches to the analysis of mixture models come with a further challenge: the likelihood function is invariant under permutation of the component labels. This leads to a phenomenon called *label switching* meaning that the labels associated with mixture components can switch during and between runs of the MCMC algorithm (Richardson and Green, 1997). In our application, some of the organelles maybe annotated with known marker proteins and this places a lower bound on the number of components, furthermore there is no label switching for these components. Bringing these ideas together results in a semi-supervised Bayesian approach, which we call Novelty TAGM.

We apply Novelty TAGM to 9 spatial proteomic datasets across a diverse range of protocols, including *hyper*LOPIT (Christoforou *et al.*, 2016; Mulvey *et al.*, 2017), LOPIT-DC (Geladaki *et al.*, 2018), dynamic organeller maps (Itzhak *et al.*, 2016) and spatial-temporal methods (Beltran *et al.*, 2016). We first validate our approach by recovering chromatin enrichment preparation in *hyper*LOPIT experiments, including mouse pluripotent stem cells and human U2OS cells. Application of Novelty TAGM to each dataset reveal additional biologically relevant compartments. Notably, we demonstrate that the U2OS *hyper*LOPIT dataset reveals 4 sub-nuclear compartments: the nucleolus, nucleoplasm, chromatin-associated, and the nuclear membrane. These findings are validated using the human protein atlas (Thul *et al.*, 2017). In addition, an endosomal enriched compartment is robustly identified across *hyper*LOPIT and LOPIT-DC technologies. We also able to uncover small collections of proteins; for example we identify vesicle proteins traffiking from the ER to the early Golgi in a *hyper*LOPIT experiment on *Saccharomyces cerevisiae*.

2 Methods

2.1 Datasets

We provide a brief description of datasets used in this manuscript. We analyse *hyper*LOPIT datasets, in which sub-cellular fractionation is performed using density-gradient centrifugation (Dunkley *et al.*, 2004, 2006; Mulvey *et al.*, 2017), on mouse pluripotent stem cells (Christoforou *et al.*, 2016) and human osteosarcoma (U2OS) cells (Thul *et al.*, 2017; Geladaki *et al.*, 2018), as well as a *Saccharomyces cerevisiae* (yeast) dataset (Nightingale *et al.*, 2019). The mouse stem cell dataset combines two 10-plex biological replicates and quantitative information on 5032 proteins. The U2OS dataset combines three 20-plex biological replicates and provides information on 4883 proteins. The yeast dataset represents four 10-plex biological replicate experiments performed in *Saccharomyces cerevisiae* cultured to early-mid exponential phase. This dataset contains quantitative information for 2846 proteins that were common across four 10-plex biological replicate experiments. Tandem Mass Tag (TMT) (Thompson *et al.*, 2003) labelling was used in all *hyper*LOPIT experiments with LC-SPS-MS3 used for high accuracy MS-based quantitation (Ting *et al.*, 2011; McAlister *et al.*, 2014). Beltran *et al.* (2016) integrated a temporal component to the analysis of the LOPIT protocol. They analysed HCMV infected primary fibroblast cells over 5 days, producing a control and infected map every 24 hours. We analyse the control and infected map 24 hours post infection, providing information on 2220 and 2196 proteins respectively.

Our approach is not limited to spatial proteomics data where the sub-cellular fractionation is performed using density gradients. We additionally analyse the dynamic organeller maps (DOM) protocols (Itzhak *et al.*, 2016, 2017), which quantify 3766 and 8985 respectively. These approaches used label-free quantitation with fractionation performed using differential centrifugation. We analyses 6 replicates from the HeLa cell line analyses in Itzhak *et al.* (2016) and 3 replicates of mouse primary neurons from Itzhak *et al.* (2017). Hirst *et al.* (2018) also used the DOM protocol and coupled CRISPR-CAS9 knockouts with spatial proteomics designed to explore the functional role of AP-5. We analyse the control map from this experiment. Finally, we analyse U2OS data using the LOPIT-DC protocol (Geladaki *et al.*, 2018), which quantified 6837 across 3 biological replicates. Though we do not analyse an PCP based spatial proteomics datasets in this manuscript, our method also applies to such data (Foster *et al.*, 2006; Kristensen *et al.*, 2012; Kristensen and Foster, 2014).

2.2 Model

2.2.1 Spatial proteomics mixture model

In this section, we briefly review the TAGM model proposed by (Crook *et al.*, 2018). Let N denote the number of observed protein profiles each of length L, corresponding to the number of quantified fractions. The quantitative profile for the i-th protein is denoted by $\mathbf{x}_i = [x_{1i}, \dots, x_{Li}]$. We suppose that there are K known sub-cellular compartments to which each protein could localise (e.g. cytoplasm, endoplasmic reticulum, mitochondria, ...). For simplicity of exposition, we refer to these K sub-cellular compartments as *components*, and introduce component labels z_i, so that $z_i = k$ if the i-th protein localises to the k-th component. To fix notation, we denote by X_L the set of proteins whose component labels are known, and by X_U the set of unlabelled proteins. If protein i is in X_U, we want the probability that $z_i = k$ for each $k = 1, \dots, K$; that is, for each unlabelled protein, the probability of belonging to each component (given a model and the observed data).

The distribution of quantitative profiles associated with each protein that localise to the k-th component is modelled as multivariate normal with mean vector $\boldsymbol{\mu}_k$ and covariance matrix Σ_k. However, many proteins are dispersed and do not fit this assumption. To model these "outliers" Crook *et al.* (2018) introduced a further indicator variable ϕ. Each protein $\mathbf{x}_i$ is then described by an additional variable ϕ_i, with $\phi_i = 1$ indicating that protein $\mathbf{x}_i$ belongs to a organelle derived component and $\phi_i = 0$ indicating that protein $\mathbf{x}_i$ is not well described by these known components. This outlier component is then modelled as a multivariate T distribution with degrees of freedom κ, mean vector $\mathbf{M}$, and scale matrix V. Thus the model can be written as:

$$\mathbf{x}_i | z_i = k, \phi_i \quad \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)^{\phi_i} \mathcal{T}(\kappa, \boldsymbol{M}, V)^{1-\phi_i}. \tag{1}$$

Let $f(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$ denote the density of the multivariate normal with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ evaluated at $\mathbf{x}$ and $g(\mathbf{x}|\kappa, \mathbf{M}, \mathbf{V})$ denote the density of the multivariate T-distribution. For any i, the prior probability of the i-th protein localising to the k-th component is denoted by $p(z_i = k) = \pi_k$. Letting $\boldsymbol{\theta} = \{\boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^K$ denote the set of all

component mean and covariance parameters, and $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^K$ denote the set of all mixture weights, it follows that:

$$p(\mathbf{x}_i|\boldsymbol{\theta}, \boldsymbol{\pi}, \phi_i, \kappa, \mathbf{M}, V) = \sum_{k=1}^{K} \pi_k \left(f(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k)^{\phi_i} g(\mathbf{x}_i|\kappa, \boldsymbol{M}, V)^{1-\phi_i} \right). \tag{2}$$

For any i, we set the prior probability of the i-th protein belonging to the outlier component as $p(\phi_i = 0) = \epsilon$.

The equation (2) can then be rewritten in the following way:

$$p(\mathbf{x}_i|\boldsymbol{\theta}, \boldsymbol{\pi}, \kappa, \epsilon, \mathbf{M}, V) = \sum_{k=1}^{K} \pi_k \left((1-\epsilon)(f(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k) + \epsilon g(\mathbf{x}_i|\kappa, \boldsymbol{M}, V)) \right), \tag{3}$$

As in Crook *et al.* (2018), we fix $\kappa = 4$, $\mathbf{M}$ as the global mean, and V as half the global variance of the data, including labelled and unlabelled proteins. To extend this model to permit novelty detection, we specify the maximum number of components $K_{max} > K$. Our proposed model then allows up to $K_{novelty} = K_{max} - K$ new phenotypes to be detected. Equation 3 can then be written as

$$p(\mathbf{x}_i|\boldsymbol{\theta}, \boldsymbol{\pi}, \kappa, \epsilon, \mathbf{M}, V) = \sum_{k=1}^{K_{max}} \pi_k \left((1-\epsilon)(f(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k) + \epsilon g(\mathbf{x}_i|\kappa, \boldsymbol{M}, V)) \right), \tag{4}$$

where the first K components correspond to known sub-cellular niches and the new phenotypes that can be inferred. The parameter sets are then augmented to include these possibly new components; that is, we redefine $\boldsymbol{\theta} = \{\boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^{K_{max}}$ to denote the set of all component mean and covariance parameters, and $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^{K_{max}}$ denote the set of all mixture weights. In practice, K_{max} is set to be large such that there is no restriction on the number of phenotypes that can be detected. Relying on the principle of over-fitted mixtures Rousseau and Mengersen (2011), components that are not supported by the data are left empty with no proteins allocated to them. We find setting $K_{novety} = 10$ is ample to detect new phenotypes.

2.2.2 Bayesian inference and convergence

We perform fully Bayesian inference, using Markov-chain Monte-Carlo methods. To be more precise, we make modifications to the collapsed Gibbs sampler approach used in Crook *et al.* (2018). The first difference is in the initialisation of the algorithm, in which half the protein are randomly selected and then randomly partitioned amongst the potential new phenotypes. Secondly, whenever a component has no proteins allocated to it the parameters are proposed from their prior. Since the number of occupied components at each iteration is random, we can monitor this quantity as a convergence diagnostic. At convergence the number of occupied components is not necessarily fixed, but oscillates around a fixed mode.

2.2.3 Label switching and post-processing

Bayesian inference in mixture models suffers form an identifiability issue known as *label switching* - a phenomenon where the allocations labels can flip during or between runs of the algorithm (Stephens, 2000). This occurs because of the symmetry of the likelihood function under permutations of these labels. We note that this only occur in unsupervised or semi-supervised mixture models. This makes inference of the parameters mixture model challenging. In our setting the labels for the known components do not switch and so standard Monte-Carlo techniques can be used to perform inference of the quantities of interest. However, for the new phenotypes label switching must occur. One standard approach to circumvent this issue is to form the so-called *posterior similarity matrix* (PSM). The PSM is an $N \times N$ matrix where the $(i,j)^{th}$ entry is the posterior probability that protein i and protein j reside in the same component. More precisely, if we let S denote the PSM and T denote the number of Monte-Carlo iterations then

$$S_{ij} = P(z_i = z_j | X, \boldsymbol{\theta}, \boldsymbol{\pi}, \kappa, \epsilon, \mathbf{M}, V) \approx \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}(z_i^{(t)} = z_j^{(t)}), \tag{5}$$

where $\mathbb{I}$ denotes the indicator function. The PSM is clearly invariant to label switching and so avoids the issues arising from the *label switching* problem. However, this immediately raises the question of how we summarise this matrix into an allocation vector. We take the approach proposed by Fritsch and Ickstadt (2009). They proposed the adjusted Rand index (AR) (Rand, 1971; Hubert and Arabie, 1985), a measure of cluster similarity, as a utility function and then we wish to find the allocation vector $\hat{z}$ that maximises the expected adjusted Rand index with respect to the true clustering z. Formally, we write

$$\hat{z} = \arg\max_{z^*} E[AR(z^*, z) | X, \boldsymbol{\theta}, \boldsymbol{\pi}, \kappa, \epsilon, \mathbf{M}, V], \tag{6}$$

which is known as the Posterior Expected Adjusted Rand index (PEAR). One obvious pitfall is that this quantity depends on the unknown true clustering z. However, this can be approximated from the MCMC samples:

$$PEAR \approx \frac{1}{T} \sum_{t=1}^{T} AR(z^*, z^{(t)}). \tag{7}$$

The space of all possible clustering over which to maximise is infeasibly large to explore. Thus we take an approach taken in Fritsch and Ickstadt (2009) to propose candidate clusterings over which to maximise. Using hierarchical clustering with distance $1 - S_{ij}$, the PEAR criterion is computed for clusterings at every level of the hierarchy. The optimal clustering $\hat{z}$ is the allocation vector which maximises the PEAR.

2.2.4 Uncertainty Quantification

We may be interested in quantifying the uncertainty in whether a protein belongs to a new sub-cellular component. Indeed it is important to distinguish whether a protein belongs to a

new phenotype or if we simply have large uncertainty about its localisation. The probability that protein i belongs to a new component is computed from the following equation:

$$P(z_i \in \{K+1, ..., K_{max}\}|X) = 1 - P(z_i \in \{1, ..., K\}|X), \tag{8}$$
$$\tag{9}$$

which we can approximate by the following Monte-Carlo average:

$$1 - \frac{1}{T}\sum_{t=1}^{T} P(z_i^{(t)} \in \{1, ..., K\}|X) = 1 - \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K} P(z_i^{(t)} = k|X) \tag{10}$$

Since the summation in equation 10 only goes up to K (the number of annotated organelles), this equation is identifiable. Throughout we refer to this as the discovery probability.

3 Results

3.1 Validating experimental design in *hyper*LOPIT

To validate Novelty TAGM we apply our method to a mouse pluripotent embryonic stem cell *hyper*LOPIT dataset (Christoforou *et al.*, 2016) and a recent human osteosarcoma (U2OS) *hyper*LOPIT dataset (Thul *et al.*, 2017; Geladaki *et al.*, 2018). The experimental protocols associated with these approaches used an additional chromatin enrichment step to resolve nuclear chromatin associated proteins from nuclear proteins not associated with the chromatin. We remove the nuclear, chromatin and ribosomal annotations from the datasets to test the ability of Novelty TAGM to recover these annotations.

For the mouse stem cell dataset, Novelty TAGM reveals 8 new phenotypes for which there is at least 1 protein with discovery probability greater than 0.95. We perform a GO enrichment analysis with FDR control perform according the Benjamini-Höchberg procedure (Benjamini and Hochberg, 1995; Ashburner *et al.*, 2000; Yu *et al.*, 2012). The proteins in each phenotype are assessed in turn and we test for enriched cellular compartment terms, against the background of all quantified protein in this experiment. Novelty TAGM recovers these hidden annotations with phenotype 2 having the enriched terms associated with chromatin, such as *chromatin* and *chromosome* ($p < 10^{80}$). Phenotype 3 corresponds to a separate nuclear substructure with enrichment for the terms *nucleolus* ($p < 10^{-60}$) and *nuclear body* ($p < 10^{-30}$). Thus, in the mouse stem cell dataset Novelty TAGM confirms the chromatin enrichment preparation designed to separate chromatin and non-chromatin associated nuclear proteins (Mulvey *et al.*, 2017). In addition, phenotype 4 demonstrates enrichment for the ribosome annotation ($p < 10^{-35}$). Phenotype 1 is enriched for *centrosome* and *microtubule* annotations ($p < 10^{-15}$), though observing the PSM in figure 1 we can see there is much uncertainty in this phenotype, giving guidance as to whether additional expert annotation should be made.

Now turning to the U2OS cancer cell-line dataset, Novelty TAGM reveals 9 new phenotypes for which there is at least 1 protein with a greater discovery probability than 0.95. These phenotypes along with the uncertainty associated with them can be visualise in figure 1. To validate the chromatin enrichment step, we first consider the confocal microscopy data provided by the Human Protein Atlas (HPA) (Thul *et al.*, 2017). The HPA data provides microscopy data on precisely the same cell-line and therefore constitutes an excellent complementary resource. This *hyper*LOPIT dataset was already shown to be in strong agreement with the microscopy data. Proteins in phenotypes $3, 4, 5$ and 8 have a nuclear related annotation at their most frequent annotation in the HPA data. Then GO enrichment analysis reveals *chromatin* and *chromosome* annotations for phenotype 3 ($p < 10^{-40}$). Phenotype 4 is enriched for the *nucleolus* ($p < 10^{-60}$), furthermore nucleoli and nucleoli/nucleus are the 2^{nd} and 3^{rd} most frequent HPA annotation for proteins belonging to this phenotype. For phenotype 5 the most associated term is nucleoplasm from the HPA data, as well as GO enrichment ($p < 10^{-10}$). Interestingly, phenotype 8 has the nuclear membrane as its most frequent HPA annotation and this is supported by GO enrichment with the terms nuclear membrane and nuclear envelope associated with proteins in this phenotype ($p < 10^{-10}$). Thus Novelty TAGM has not only confirmed successful validation of chromatin enrichment but also demonstrated other sub-nuclear level resolution. Phenotype 1 is enriched for *ribo-*

some ($p < 10^{-20}$), whilst phenotype 2 is enriched for *endosomes* ($p < 10^{-30}$). Unfortunately, there are no HPA annotation for ribosome or endosome to provide additional validation.

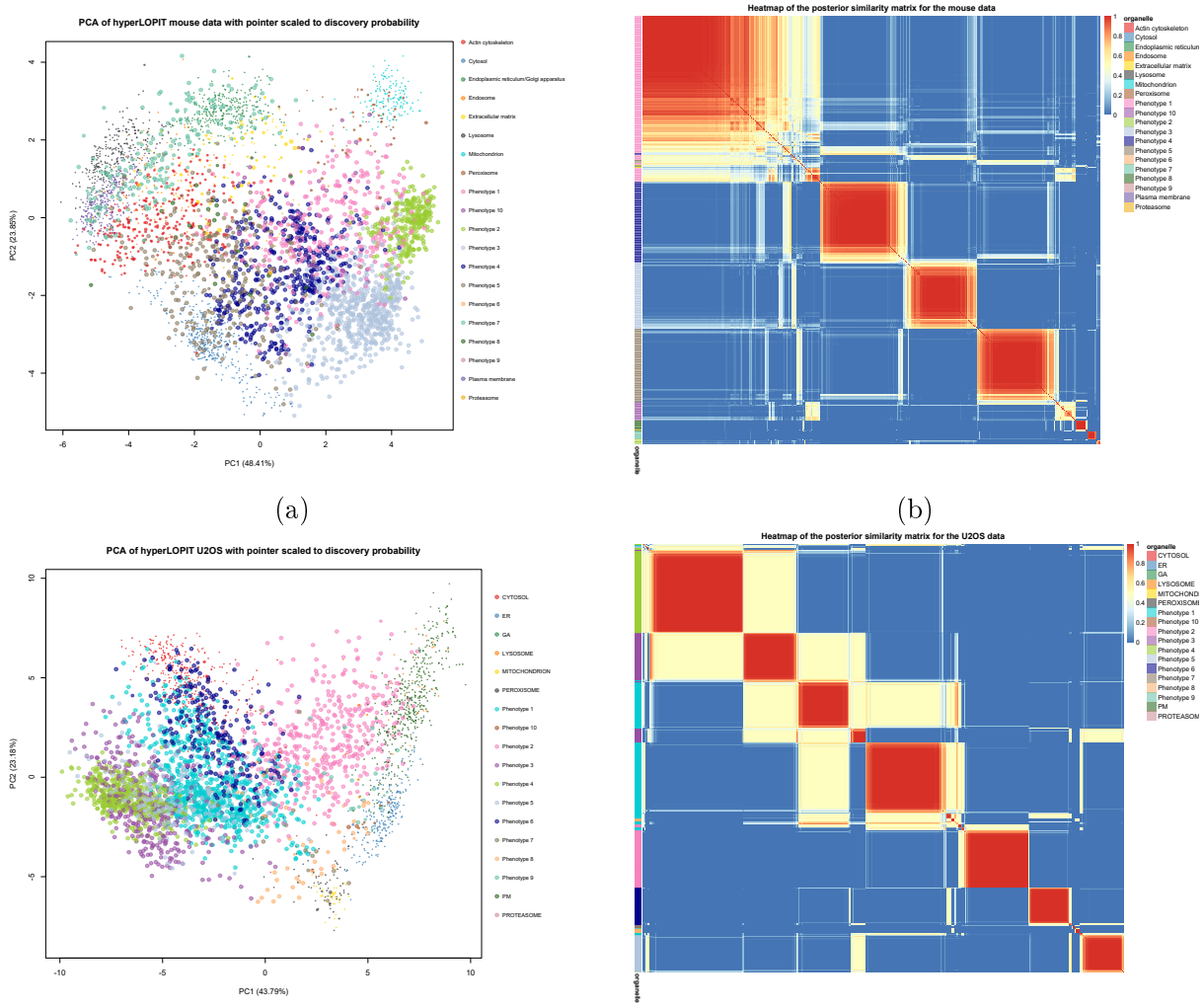


Figure 1: (a) A PCA plot of the *hyper*LOPIT mouse pluripotent stem cell data. The points are coloured according the organelle or proposed new phenotype and are scaled according the the discovery probability. The PCA plot reveals clear clustering structure in the data and confidently identified new phenotypes. (b) A heatmap of the posterior similarity matrix derived from the mouse stem cell data demonstrating the uncertainty in the clustering structure of the data. We have only plotted the proteins who have greater than 0.99 probability of belonging to a new phenotype and probability of being an outlier less than 0.95. (c) A PCA plot of the *hyper*LOPIT U2OS cancer cell-line data. The points are coloured according the organelle or proposed new phenotype and are scaled according the the discovery probability. The PCA plot reveals clear clustering structure in the data and confidently identified new phenotypes. (d) A heatmap of the posterior similarity matrix derived from the U2OS cell line data demonstrating the uncertainty in the clustering structure of the data. We have only plotted the proteins who have greater than 0.99 probability of belonging to a new phenotype and probability of being an outlier less than 0.95.

3.2 Uncovering additional sub-cellular structures

Having validated the ability of Novelty TAGM to recover known experimental design, as well as uncover additional sub-cellular niches resolved in the data we turn to apply Novelty TAGM to several additional dataset. We first consider the LOPIT-DC dataset on the U2OS cell line (Geladaki *et al.*, 2018). For additional validation of our proposed method we removed the nuclear, proteosomal, and ribosomal annotations. Novelty TAGM reveals 10 phenotypes with at least 1 protein with a discovery probability of greater than 0.99 and outlier probability of less than 0.95. These clusters and the uncertainty associated with them can be visualised in figure 2

In a similar vein to the analysis performed on the *hyper*LOPIT U2OS dataset, we initially use the available HPA to validate these clusters (Thul *et al.*, 2017). Phenotypes $3, 5, 7$ and 9 have their most frequent HPA annotation as nuclear associated, with at least 10% of the proteins in these phenotypes with this annotation. To obtain additional functional information about these phenotypes, we perform a over-representation analysis on GO cellular compartment terms. Phenotype 3 reveals both nucleolus ($p < 10^{-60}$) and ribosome ($p < 10^{-30}$) annotations. Phenotype 5 reveals a proteasome cluster ($p < 10^{-30}$). A chromatin enriched phenotype is also discovered, with phenotype 9 having chromosome ($p < 10^{-60}$) and chromatin ($p < 10^{-40}$) terms significantly over-represent in these clusters. Phenotype 6 represents a cluster with mixed annotation with over-representation for both plasma membrane ($p < 10^{-8}$) and the extracellular matrix ($p < 10^{-2}$), this is supported by HPA annotation with vesicles, cytosol, and plasma membrane the top three annotations. Furthermore, phenotype 8 is significantly enriched for endosomes ($p < 10^{-55}$). In addition, 107 of the proteins in this phenotype are also localised to the endosome-enriched phenotype presented in the U2OS *hyper*LOPIT dataset. Thus, we robustly identify new phenotypes across highly different spatial proteomics protocols. Hence, we have presented strong evidence for additional annotations in this dataset beyond the original analysis of this dataset (Geladaki *et al.*, 2018); in particular, we have described sub-nuclear resolution with separated chromatin and non-chromatin classes. Furthermore, we have joint evidence for an endosomal cluster in both the LOPIT-DC and *hyper*LOPIT dataset. Furthermore, we have quantified uncertainty in these proposed phenotypes providing rich information for rigorous interrogation of these datasets.

Novelty TAGM uncovers 8 phenotypes in the yeast *hyper*LOPIT data with at least 1 protein with discovery probability greater than 0.95. 4 of these phenotypes have no significant over-represented annotations. The first phenotype is enriched for the cell periphery ($p < 10^{-19}$) and fungal-type vacuole ($p < 10^{-10}$). Phenotype 3 has over-represented annotations for the kinetochore ($p < 0.01$), whilst phenotype 4 is enriched for the cytoskeleton ($p < 10^{-7}$). Phenotype 4 represent a joint golgi and ER cluster with the COPII-coated ER to Golgi transport vesicle enriched in this phenotype ($p < 10^{-14}$), along with the endoplasmic reticulum membrane ($p < 10^{-10}$) and the Golgi membrane ($p < 10^{-9}$). This phenotype is a collection of proteins with the role of trafficking material from the ER to the early Golgi.

We apply Novelty TAGM to the HCMV infected fibroblast cells 24 hours post infection (hpi) (Beltran *et al.*, 2016), and discover 9 additional phenotypes with at least 1 protein with discovery probability greater than 0.95 demonstrated in figure 2. Phenotype 2 contains a singleton protein and phenotypes $4, 6, 7, 8$ and 9 are not significantly enriched for any

annotations. However, phenotype 3 is enriched for the mitochondrial membrane and envelope annotations ($p < 10^{-4}$), this is an addition to the already annotated mitochondrial class indicating sub-mitochondrial resolution. Phenotype 1 is a mixed ribosomal/nuclear cluster with enrichment for nucleoplasm ($p < 10^{-5}$) and the small ribosomal subunit ($p < 10^{-4}$), which is distinct from phenotype 5 which is enriched for the large ribosomal subunit ($p < 10^{-10}$). This demonstrates unbiased separation of the two ribosomal subunits, which was overlooked in the original analysis (Beltran *et al.*, 2016).

Novelty TAGM reveals 7 phenotypes with at least 1 protein with discovery probability greater than 0.95. Phenotypes $4, 6$ and 9 have no significantly enriched Gene Ontology terms (threshold $p = 0.01$). However, we observe 2 phenotypes with ribosomal enrichment with the large ribosomal subunit shared across phenotypes 1 and 5 with significance at levels $p < 10^{-4}$ and $p < 10^{-7}$, respectively. We observer significant enrichment for the actin cytoskelton ($p < 10^{-5}$) in phenotype 2. Phenotype 3 represents a mixed peroxisome ($p < 10^{-3}$) and mitchondrion cluster ($p < 10^{-3}$), an unsurprising results since the mitchondrion and peroxisome have similar biochemical fractionation properties. The differing number of confidently identified and biologically relevant phenotypes discovered between the two fibroblast datasets, could be down to the differing levels of structure between the two datasets. From figure 3 we see differing levels of clustering structure in these datasets.

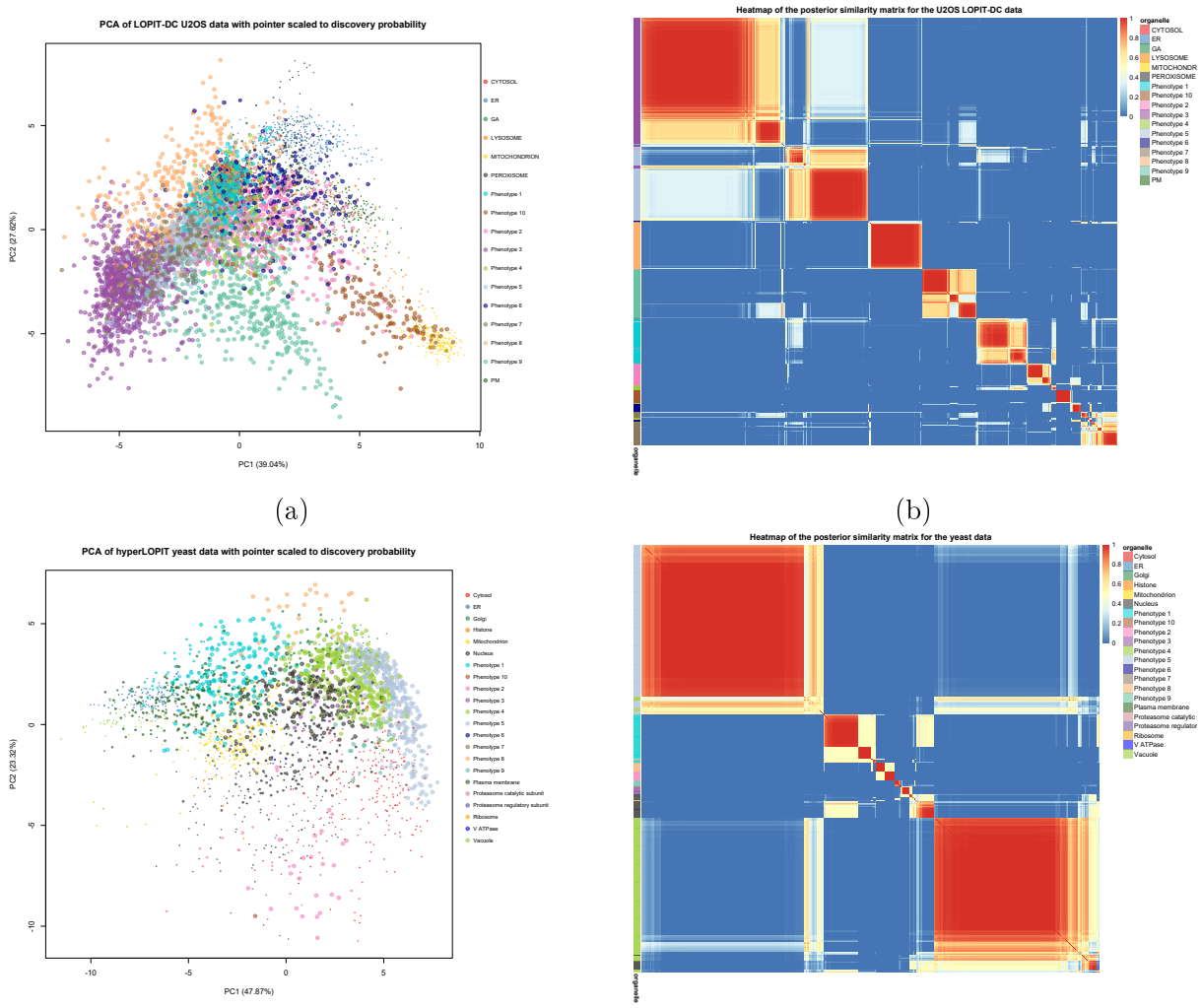


Figure 2: (a) A PCA plot of the LOPIT-DC U2OS data. The points are coloured according to the organelle or proposed new phenotype and are scaled according the the discovery probability. The PCA plot reveals clear clustering structure in the data and confidently identified new phenotypes. (b) A heatmap of the posterior similarity matrix derived from the U2OS demonstrating the uncertainty in the clustering structure of the data. We have only plotted the proteins which have greater than 0.99 probability of belonging to a new phenotype and probability of being an outlier less than 0.95. (c) A PCA plot of the *hyper*LOPIT yeast data. The points are coloured according the organelle or proposed new phenotype and are scaled according the the discovery probability. The PCA plot reveals clear clustering structure in the data and confidently identified new phenotypes. (d) A heatmap of the posterior similarity matrix derived from the yeast data demonstrating the uncertainty in the clustering structure of the data. We have only plotted the proteins who have greater than 0.99 probability of belonging to a new phenotype and probability of being an outlier less than 0.95

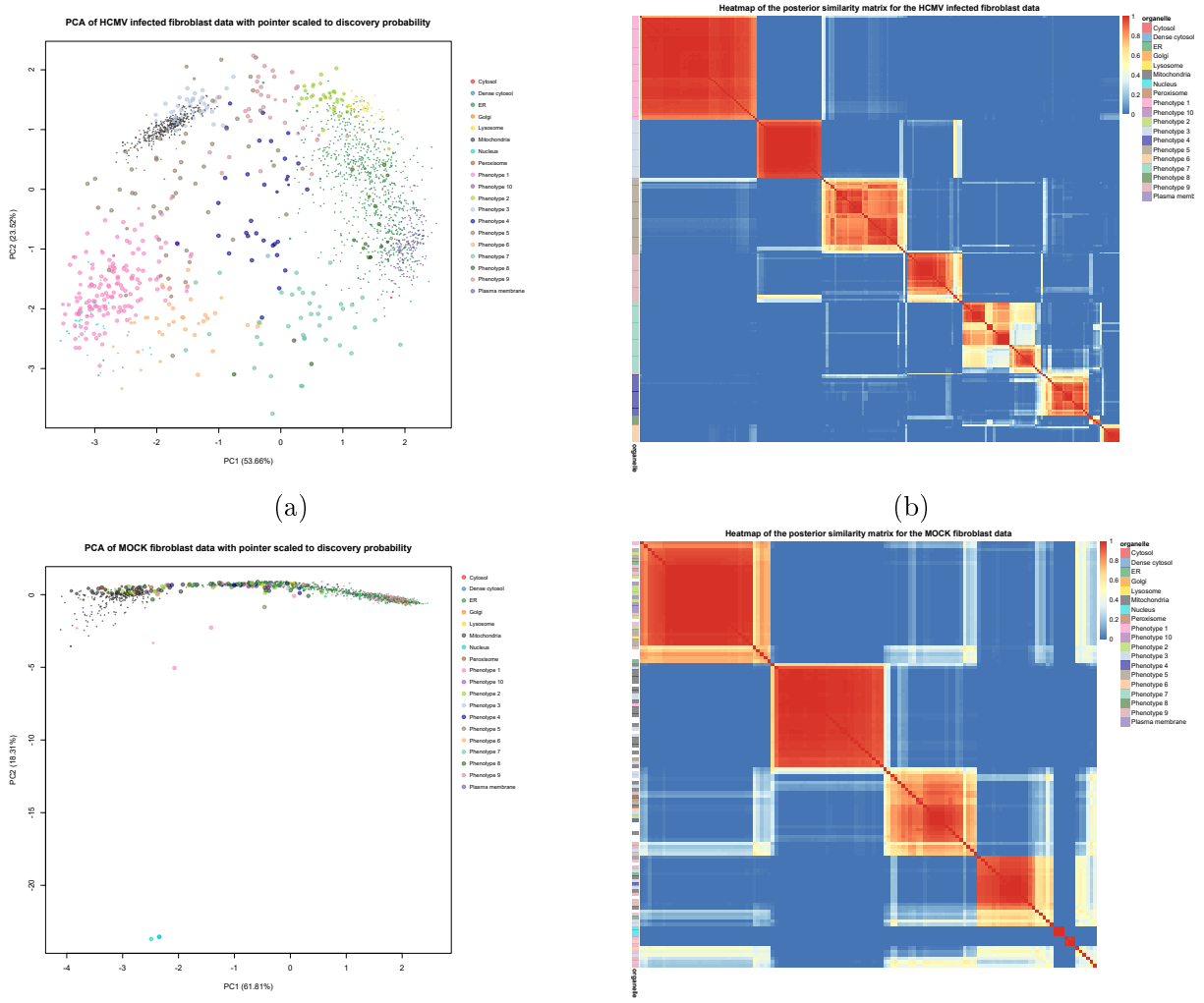


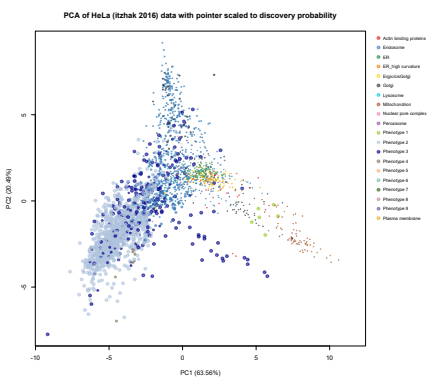
Figure 3: (a) A PCA plot of the HCMV infected fibroblast data 24 hpi. The points are coloured according to the organelle or proposed new phenotype and are scaled according the the discovery probability. The PCA plot reveals clear clustering structure in the data and confidently identified new phenotypes. (b) A heatmap of the posterior similarity matrix derived from the fibroblast data demonstrating the uncertainty in the clustering structure of the data. We have only plotted the proteins which have greater than 0.99 probability of belonging to a new phenotype and probability of being an outlier less than 0.95. (c) A PCA plot of the mock fibroblast data 24 hpi. The points are coloured according to the organelle or proposed new phenotype and are scaled according the the discovery probability. The PCA plot reveals clear clustering structure in the data and confidently identified new phenotypes. (d) A heatmap of the posterior similarity matrix derived from the mock fibroblast data demonstrating the uncertainty in the clustering structure of the data. We have only plotted the proteins who have greater than 0.99 probability of belonging to a new phenotype and probability of being an outlier less than 0.95

3.3 Refining annotation in organeller maps

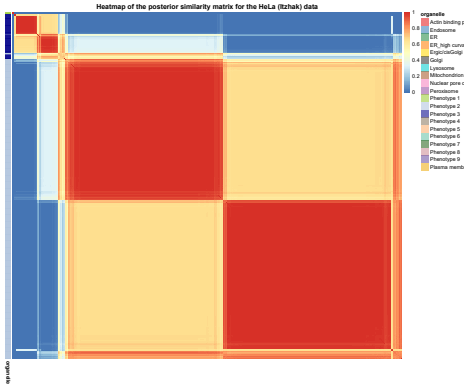
The dynamic organeller map (DOM) protocol was developed to reduce the time taken to perform MS-based spatial proteomic mapping albeit at the cost of organelle resolution (Itzhak *et al.*, 2016; Gatto *et al.*, 2018). The three dataset analysed here are two HeLa cell line (Itzhak *et al.*, 2016; Hirst *et al.*, 2018) and a mouse primary neuron dataset (Itzhak *et al.*, 2017). All three of these datasets have been annotated to contain a mixed class called "Large Protein Complexes". This class likely contains a mixture of cytosolic, ribosomal and nuclear sub-compartment that pellet during the centrifugation step used to capture this mixed fraction. We apply Novelty TAGM to these data and remove this "Large Protein Complex" annotation, to derive more precise annotations for these datasets. The HeLa dataset of Itzhak *et al.* (2016), which we refer to as HeLa Itzhak, has 3 additional phenotypes uncovered by Novelty TAGM. The first phenotype is enriched for the mitochondrial membrane ($p < 0.01$) distinct from the already annotated mitochondrial class. Phenotype 2 represent a mixed cluster with nuclear, ribosomal and cytosolic enriched terms such as cytosolic ribosome ($p < 10^{-40}$), nucleolus ($p < 10^{-30}$) and cytosolic part ($p < 10^{-25}$). The final phenotype is enriched for chromatin and chromosome ($p < 10^{-10}$) suggesting sub-nuclear resolution. Furthermore, as a result of quantifying uncertainty, we can see in figure 4 there are potentially more sub-cellular structures. However, the uncertainty is too great to support these phenotypes.

The mouse neuron dataset reveals 10 phenotypes after we apply Novelty TAGM. However 8 of these phenotypes have no biological relevance. This is likely a manifestation of the dispersed nature of this dataset, where the variability is generated by technical artefacts rather than biological signal. However, despite this Novelty TAGM is able to detect two relevant phenotypes: the first phenotype is enriched for nucleolus ($p < 0.01$); the second for chromosome ($p < 0.01$). This is suggests additional annotations for this dataset.

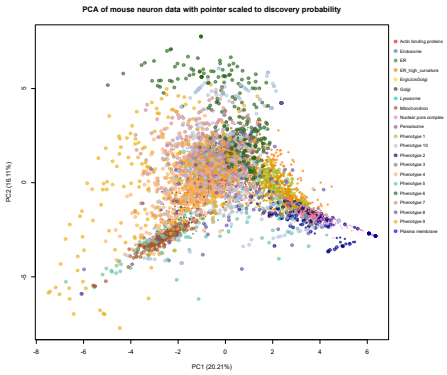
The HeLa dataset of Hirst *et al.* (2018), which we refer to HeLa Hirst, reveals 7 phenotypes with at least 1 protein with discovery probability greater than 0.95. However, three of these phenotypes represent singleton proteins. Phenotype 1 reveals a mixed cytosol/ribosomal annotations with the terms cytosolic ribosome ($p < 10^{-30}$) and cytosolic part ($p < 10^{-25}$) significantly over-represented. There are no further phenotypes enriched annotations (threshold $p = 0.01$), except phenotype 2 which is represents a mixed extracellular and cytosolic cluster. For example the terms extracellular organelle ($p < 10^{-13}$) and cytosol ($p < 10^{-10}$) are over-represented.



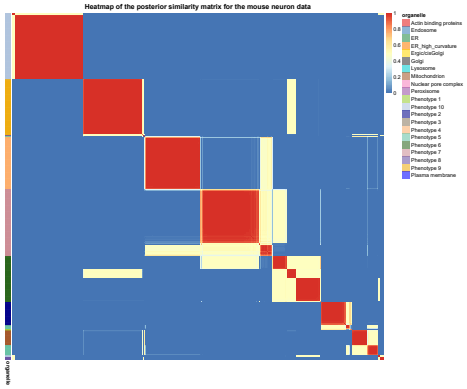
(a)



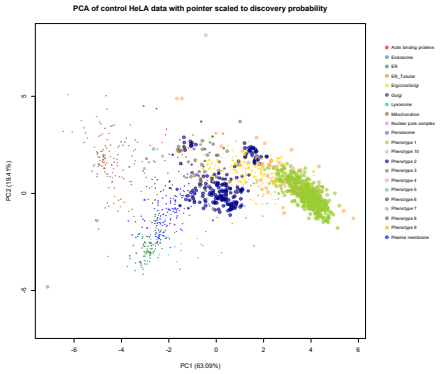
(b)



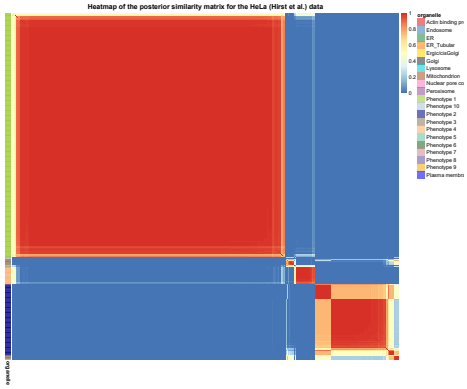
(c)



(d)



(e)



(f)

Figure 4: (a),(b),(c) PCA plots of the HeLa Itzhak data, mouse neuron data and HeLa Hirst data. The pointers are coloured according to the assigned organelle or phenotype and scaled according to their discovery probability. (d),(e),(f) Heatmaps of the HeLa Itzhak data, mouse neuron data and HeLa Hirst data. Only the proteins whos discovery probability is greater than 0.99 and outlier probability less than 0.95 are shown. The heatmaps demonstrate the uncertainty in the clustering structure present in the data.

4 Discussion

We have presented a semi-supervised Bayesian approach that simultaneously allows probabilistic allocation of proteins to organelles, detection of outlier proteins, as well as the discovery of sub-cellular structures. Our method unifies several approaches present in the literature, marrying the ideas of supervised machine learning and unsupervised structure discovery. Formulating inference in a Bayesian framework allows for the quantification of uncertainty; in particular, the uncertainty in the number of newly discovered annotations.

Our proposed methodology allows us to interrogate individual proteins to see whether they belong to a newly discovered phenotype. Through the posterior similarity matrix we can visualise the global patterns in the uncertainty in phenotype discovery. We summarise this posterior similarity matrix into a single clustering by maximising the posterior expected adjusted rand index. This methodology infers the number of clusters supported by the data, rather than many ad-hoc approaches which require specification of the number of clusters.

Application of our method across 9 different spatial proteomics experiments with diverse protocols and varying levels of resolution revealed additional annotation in every single experiment. Our analysis recovered and validated chromatin enrichment preparation experimental design in *hyper*LOPIT datasets. Our approach also revealed additional sub-cellular niches in the mouse stem cell dataset and U2OS *hyper*LOPIT dataset.

Our method revealed resolution of 4 sub-nuclear compartments in the U2OS *hyper*LOPIT dataset, which was validated by human protein atlas annotations. An additional endosome-enriched phenotype was uncovered and Novelty TAGM robustly identified an overlapping phenotype in LOPIT-DC data providing strong evidence for endosomal resolution. Further biological relevant annotations were uncovered in these datasets as well as other datasets. For example, a group of vesicle proteins involved in transport from the ER to the early Golgi was identified in the yeast dataset; resolution of the ribosomal subunit was identfied in the fibroblast dataset, and seperate nuclear, cytosolic and ribosomal annotations were identified in DOM datasets.

Thus our approach is widely applicable within the field of spatial proteomics and builds upon state-of-the-art approaches. The computational algorithms presented here are disseminated as part of the Bioconductor project (Gentleman *et al.*, 2004; Huber *et al.*, 2015) building on MS-based data structures provided in Gatto and Lilley (2012) and are disseminated as part of the pRoloc, with all data provided in pRolocdata (Gatto *et al.*, 2014b).

During our analysis, we observed that the posterior similarity matrices have potential sub-clustering structures. Many known organelles and sub-cellular niches have sub-compartmentalisation, thus methodology to detect sub-compartmentalisation is in preparation. Furthermore, we have observed that different experiments and different data modalities share information. Integrative approach to spatial proteomics analysis are also desired.

References

Antoniak, C. E. (1974). Mixtures of dirichlet processes with applications to bayesian non-parametric problems. *Ann. Statist.*, **2**(6), 1152–1174.

Ashburner, M. et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, **25**(1), 25–29.

Beltran, P. M. J. et al. (2016). A portrait of the human organelle proteome in space and time during cytomegalovirus infection. *Cell systems*, **3**(4), 361–373.

Benjamini, Y. et al. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300.

Breckels, L. M. et al. (2013). The effect of organelle discovery upon sub-cellular protein localisation. *Journal of proteomics*, **88**, 129–140.

Christoforou, A. et al. (2016). A draft map of the mouse pluripotent stem cell spatial proteome. *Nature communications*, **7**, 9992.

Crook, O. M. et al. (2018). A bayesian mixture modelling approach for spatial proteomics. *bioRxiv*.

Dunkley, T. P. et al. (2004). Localization of organelle proteins by isotope tagging (lopit). *Molecular & Cellular Proteomics*, **3**(11), 1128–1134.

Dunkley, T. P. et al. (2006). Mapping the arabidopsis organelle proteome. *Proceedings of the National Academy of Sciences*, **103**(17), 6518–6523.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.*, **2**(4), 615–629.

Foster, L. J. et al. (2006). A mammalian organelle map by protein correlation profiling. *Cell*, **125**(1), 187–199.

Fritsch, A. et al. (2009). Improved criteria for clustering based on the posterior similarity matrix. *Bayesian Anal.*, **4**(2), 367–391.

Gatto, L. et al. (2012). Msnbase - an r/bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. *Bioinformatics*, **28**, 288–289.

Gatto, L. et al. (2010). Organelle proteomics experimental designs and analysis. *Proteomics*, **10**(22), 3957–3969.

Gatto, L. et al. (2014a). A foundation for reliable spatial proteomics data analysis. *Molecular & Cellular Proteomics*, pages mcp–M113.

Gatto, L. et al. (2014b). Mass-spectrometry based spatial proteomics data analysis using proloc and prolocdata. *Bioinformatics*.

Gatto, L. et al. (2018). Assessing sub-cellular resolution in spatial proteomics experiments. *bioRxiv*.

Geladaki, A. et al. (2018). Lopit-dc: A simpler approach to high-resolution spatial proteomics. *bioRxiv*.

Gentleman, R. C. et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, **5**(10), R80.

Gibson, T. J. (2009). Cell regulation: determined to signal discrete cooperation. *Trends in biochemical sciences*, **34**(10), 471–482.

Groen, A. J. et al. (2014). Identification of trans-golgi network proteins in arabidopsis thaliana root tissue. *Journal of proteome research*, **13**(2), 763–776.

Hirst, J. et al. (2018). Role of the ap-5 adaptor protein complex in late endosome-to-golgi retrieval. *PLoS biology*, **16**(1), e2004411.

Huber, W. et al. (2015). Orchestrating high-throughput genomic analysis with bioconductor. *Nature methods*, **12**(2), 115–121.

Hubert, L. et al. (1985). Comparing partitions. *Journal of Classification*, **2**(1), 193–218.

Itzhak, D. N. et al. (2016). Global, quantitative and dynamic mapping of protein subcellular localization. *Elife*, **5**, e16950.

Itzhak, D. N. et al. (2017). A mass spectrometry-based approach for mapping protein subcellular localization reveals the spatial proteome of mouse primary neurons. *Cell reports*, **20**(11), 2706–2718.

Kau, T. R. et al. (2004). Nuclear transport and cancer: from mechanism to intervention. *Nature Reviews Cancer*, **4**(2), 106–117.

Kristensen, A. R. et al. (2014). Protein correlation profiling-silac to study protein-protein interactions. In *Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC)*, pages 263–270. Springer.

Kristensen, A. R. et al. (2012). A high-throughput approach for measuring temporal changes in the interactome. *Nature methods*, **9**(9), 907.

Laurila, K. et al. (2009). Prediction of disease-related mutations affecting protein localization. *BMC genomics*, **10**(1), 122.

McAlister, G. C. et al. (2014). Multinotch ms3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Analytical chemistry*, **86**(14), 7150–7158.

Mulvey, C. M. et al. (2017). Using hyperLOPIT to perform high-resolution mapping of the spatial proteome. *Nature Protocols*, **12**(6), 1110–1135.

Nightingale, D. J. H. et al. (2019). The subcellular organisation of saccharomyces cerevisiae. *Current Opinion in Chemical Biology*, **48**(11), 1–10.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, **66**(336), 846–850.

Richardson, S. et al. (1997). On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, **59**(4), 731–792.

Rousseau, J. et al. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**(5), 689–710.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, **6**(2), 461–464.

Siljee, J. E. et al. (2018). Subcellular localization of MC4R with ADCY3 at neuronal primary cilia underlies a common pathway for genetic predisposition to obesity. *Nat Genet*.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **62**(4), 795–809.

Tan, D. J. et al. (2009). Mapping organelle proteins and protein complexes in drosophila melanogaster. *Journal of proteome research*, **8**(6), 2667–2678.

Thompson, A. et al. (2003). Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by ms/ms. *Analytical chemistry*, **75**(8), 1895–1904.

Thul, P. J. et al. (2017). A subcellular map of the human proteome. *Science*, **356**(6340), eaal3321.

Ting, L. et al. (2011). Ms3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nature methods*, **8**(11), 937.

Yu, G. et al. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology*, **16**(5), 284–287.