Inferring differential subcellular localisation in comparative spatial proteomics using BANDLE

Oliver M. Crook *[1,2,3], Colin T. R. Davies[1,3,6], Laurent Gatto[4], Paul D.W. Kirk[2,5], and Kathryn S. Lilley † [1,3]

[1] *Cambridge Centre for Proteomics, Department of Biochemistry, University of Cambridge, Cambridge, UK*
[2] *MRC Biostatistics Unit, School of Clinical Medicine, University of Cambridge, Cambridge, UK*
[3] *Milner Therapeutics Institute, Jeffrey Cheah Biomedical Centre, University of Cambridge, Cambridge, CB2 0AW, UK*
[4] *de Duve Institute, UCLouvain, Avenue Hippocrate 75, 1200 Brussels, Belgium*
[5] *Cambridge Institute of Therapeutic Immunology & Infectious Disease (CITIID), Jeffrey Cheah Biomedical Centre, Cambridge Biomedical Campus, University of Cambridge, UK.*
[6] *Mechanistic Biology and Profiling, Discovery Sciences, R&D, AstraZeneca, Cambridge, UK*

January 16, 2021

Abstract

The steady-state localisation of proteins provides vital insight into their function. These localisations are context specific with proteins translocating between different sub-cellular niches upon perturbation of the subcellular environment. *Differential localisation* provides a step towards mechanistic insight of subcellular protein dynamics. Aberrant localisation has been implicated in a number of pathologies, thus *differential localisation* may help characterise disease states and facilitate rational drug discovery by suggesting novel targets. High-accuracy high-throughput mass spectrometry-based methods now exist to map the steady-state localisation and re-localisation of proteins. Here, we propose a principled Bayesian approach, BANDLE, that uses these data to compute the probability that a protein differentially localises upon cellular perturbation, as well quantifying the uncertainty in these estimates. Furthermore, BANDLE allows information to be shared across spatial proteomics datasets to improve statistical power. Extensive simulation studies demonstrate that BANDLE reduces the number of both type I and type II errors compared to existing approaches. Application of BANDLE to datasets studying EGF stimulation and AP-4 dependent localisation recovers well studied translocations, using only two-thirds of the provided data. Moreover, we implicate TMEM199 with AP-4 dependent localisation. In an application to cytomegalovirus infection, we obtain novel insights into the rewiring of the host proteome. Integration of high-throughput transcriptomic and proteomic data, along with degradation assays, acetylation experiments and a cytomegalovirus interactome allows us to provide the functional context of these data.

*omc25@cam.ac.uk, oliver.crook@stats.ox.ac.uk
†ksl23@cam.ac.uk

1 Introduction

The cell is compartmentalised into organelles and sub-cellular niches, allowing many biological process to occur in synchrony (Gibson, 2009). Proteins are localised to these niches in accordance to their function and thus to shed light on the function of a protein, it is necessary to determine its subcellular location. A number of pathologies have implicated incorrect localisation as a contributing factor, including obesity (Siljee *et al.*, 2018), cancers (Kau *et al.*, 2004), neurological disorders (Davies *et al.*, 2018), as well as multiple others (Laurila and Vihinen, 2009). It is estimated that up to 50% of proteins reside in multiple locations (Christoforou *et al.*, 2016; Thul *et al.*, 2017), which complicates the study of their localisations. Community approaches have led to substantial improvements in our understanding of sub-cellular localisation (Thul *et al.*, 2017; Sullivan *et al.*, 2018). However, image-based approaches are often low in throughput and high-throughput alternatives are desirable. Furthermore, many biological processes are regulated by re-localisation of proteins, such as transcription factors shuttling from the cytoplasm to the nucleus, which are difficult to map using imaging methods at scale (Plotnikov *et al.*, 2011).

To simultaneously study the steady-state localisation and re-localisation of proteins, one approach is to couple gentle cell lysis and cell fractionation with high-accuracy mass spectrometry (MS) (Christoforou *et al.*, 2016; Mulvey *et al.*, 2017; Geladaki *et al.*, 2019; Orre *et al.*, 2019). These approaches have already led to high-resolution maps of mouse embryonic stem cell (mESC) (Christoforou *et al.*, 2016), human cell lines (Geladaki *et al.*, 2019; Orre *et al.*, 2019), *S. cerevisiae* (bakers' yeast) (Nightingale *et al.*, 2019), cyanobacterium (Baers *et al.*, 2019) and the apicomplexan *Toxoplasma Gondii* (Barylyuk *et al.*, 2020). Dynamic experiments have given us unprecedented insight into HCMV infection (Beltran *et al.*, 2016), EGF stimulation (Itzhak *et al.*, 2016), EGFR inhibition (Orre *et al.*, 2019). In addition, CRISPR-Cas9 knockouts coupled with spatial proteomics has given insights into AP-4 vesicles (Davies *et al.*, 2018), as well as AP-5 cargo (Hirst *et al.*, 2018). In a study by Shin *et al.* (2020), the golgin long coiled-coil proteins that selectively capture vesicles destined for the Golgi were re-located to the mitochondria by replacing their Golgi targeting domains with a mitochondrial transmembrane domain (Shin *et al.*, 2020). This allowed the authors to readily observe the vesicle cargo and regulatory proteins that are redirected to the mitochondria, whilst avoiding technical issues that arise because of the redundancy of the golgins and their transient interaction with vesicles. Together, these collections of experiments suggest spatial proteomics can provide unprecedented insight into biological function.

Mass spectrometry-based spatial proteomics currently relies on supervised machine learning methods, such as *support vector machines*, to assign proteins to a sub-cellular niche using marker proteins with known localisations (Gatto *et al.*, 2010, 2014a). Advanced computational approaches have also been to developed, including novelty detection algorithms (Breckels *et al.*, 2013; Crook *et al.*, 2020a) and transfer learning approaches (Breckels *et al.*, 2016b). These approaches are implemented in the pRoloc software suite (Gatto *et al.*, 2014b; Breckels *et al.*, 2016a), which builds on the MSnbase software (Gatto and Lilley, 2012) as part of the Bioconductor project (Gentleman *et al.*, 2004; Huber *et al.*, 2015). However, most machine learning methods fail to quantify uncertainty in the assignment of a protein to an organelle, which is paramount to obtaining a rich interrogation of the data. Crook *et al.* (2018) developed a Bayesian model to analyse spatial proteomics data and highlighted that uncertainty quantification can give insights into patterns of multi-localisation. The method is implemented as a tool as part of the Bioconductor project (Crook *et al.*, 2019a).

In dynamic and comparative experiments; that is, those where we expect re-localisation upon some stimulus to sub-cellular environment, the data analysis is more challenging. The task can no longer be phrased as a supervised learning problem, but the question under consideration is clear: which proteins have different sub-cellular niches after cellular perturbation? Procedures to answer this question have been presented by authors (Beltran *et al.*, 2016; Itzhak *et al.*, 2016, 2017; Kennedy *et al.*, 2020) and reviewed in Crook *et al.* (2020b). The approach of Itzhak *et al.* (2016, 2017) relies on coupling a multivariate outlier test and a reproducibility score - termed the (movement-reproducibility) MR method. A threshold is then applied to these scores to obtain a list of proteins that re-locate; "moving" proteins. However, these scores can be challenging to interpret, since their ranges differ from one experiment to another and require additional replicates to calibrate the scores. Furthermore, the test ignores the spatial context of each protein, rendering the approach inefficient with some applications allowing false discovery rates of up to 23% (Hirst *et al.*, 2018). Finally, the approach does not quantify uncertainty which is of clear importance when absolute purification of sub-cellular niches is impossible and multi-localising proteins are present. Recently, Kennedy *et al.* (2020) introduced a computational pipeline for analysing dynamic spatial proteomics experiments by reframing it as a classification task. However, this formulation ignores that some changes in localisation might be shifts in multi-localisation patterns or only partial changes. Furthermore, their approach cannot be applied to replicated experiments and so its applicability is limited. In addition, the authors found that they needed to combine several of the organelle classes together to obtain good results. Finally, the framing of the problem as a classification task only allows a descriptive analysis of the data. These considerations motivate the development of a more sophisticated and reasoned methodology.

We present Bayesian ANalysis of Differential Localisation Experiments (BANDLE) - an integrative semi-supervised functional mixture model, to obtain the probability of a protein being differentially localised between two conditions. Posterior Bayesian computations are performed using Markov-chain Monte-Carlo and so uncertainty estimates are also available (Gilks *et al.*, 1995). We associate the term *differentially localised* to those proteins which are assigned different sub-cellular localisations between two conditions. Then, we refer precisely to this phenomenon as *differential localisation*, throughout the text. Hence, our main quantity of interest is the probability that a protein is differentially localised between two conditions.

BANDLE models the quantitative protein profiles of each sub-cellular niche in each replicate of each experiment non-parametrically (Crook *et al.*, 2019b). A first layer of integration combines replicate information in each experiment to obtain the localisation of proteins within a single experimental condition. Then a joint prior distribution on protein allocations across experimental conditions allows information to be shared across experiments and a differential localisation probability to be obtained. Two prior distributions are proposed: one using a matrix extension of the Dirichlet Distribution and another, more flexible prior, based on Pólya-Gamma augmentation (Polson *et al.*, 2013; Choi *et al.*, 2013; Linderman *et al.*, 2015).

A number of integrative mixture models have been proposed including Multiple Dataset Integration (Kirk *et al.*, 2012), infinite tensor factorisation approaches (Banerjee *et al.*, 2013), Bayesian Consensus Clustering (Lock and Dunson, 2013) and Clusternomics (Gabasova *et al.*, 2017). The methods have been developed mostly in the context of cancer sub-typing or transcriptional module discovery. Our approach is most similar to Clusternomics, which places a prior on the tensor product

between the mixing proportions; but instead our model defines mixing proportions across datasets - upon which we introduce a prior. Importantly, our approach demands that there is an explicit link between components in each dataset, which can be difficult to assume outside the semi-supervised setting because of a statistical issue known as label-switching (Richardson and Green, 1997).

To demonstrate the utility of BANDLE, we first perform extensive simulations and compare to the MR approach. Our simulation study shows that our approach reduces the number of Type I and Type II errors, and, as a result, can report an increased number of differentially localised proteins. These simulations also highlight the robustness of our approach to a number of experimental scenarios including batch effects. Our simulation studies also highlight that BANDLE provides interpretative improvements and clearer visualisations, and makes less restrictive statistical assumptions. We then apply our method to a number of datasets with well studied examples of differential localisation, including EGF stimulation and AP-4 dependent localisation. We recover known biology and provided additional cases of differential localisation, and suggest that TMEM199 localisation is potentially AP-4 dependent. Finally, we apply BANDLE to a human cytomegalovirus (HCMV) dataset - a case where MR approach is not applicable because the MR approach requires multiple replicates. Integration of high-throughput transcriptomic and proteomic data, along with degradation assays, acetylation experiments and a cytomegalovirus interactome allows us to provide the functional context of these data. In particular, we provide the spatial context of the interactome data.

2 Results

2.1 The BANDLE workflow

To perform statistical analysis of differential subcellular localisation experiments we developed BANDLE. BANDLE is a semi-supervised integrative functional mixture model that allows the computation of a differential localisation probability. The BANDLE workflow, visualised in figure 1, begins with a well defined mass-spectrometry based spatial proteomics experiment. A cellular perturbation of interest is performed alongside control experiments in wild-type cells. The usual principles of experimental design for proteomics apply, to avoid confounding (Gatto *et al.*, 2010, 2014a). Additional quality control steps are undertaken specifically for spatial proteomics experiments (Gatto *et al.*, 2010, 2014a, 2019). To apply the Bayesian model, we first calibrate the prior based on *prior predictive checks* (Gelman *et al.*, 1995). In all scenarios, we check the prior expected number of differentially localised proteins and the probability that more than l proteins are differentially localised. These are reported in the supplement. We then proceed with Bayesian parameter inference using Markov-Chain Monte-Carlo (MCMC) (Gilks *et al.*, 1995) and the checking of convergence. We visualise our results principally using rank plots, where proteins are ranked from those most likely to be differentially localised or not. Results are then interpreted using other functional screens, assays and databases.

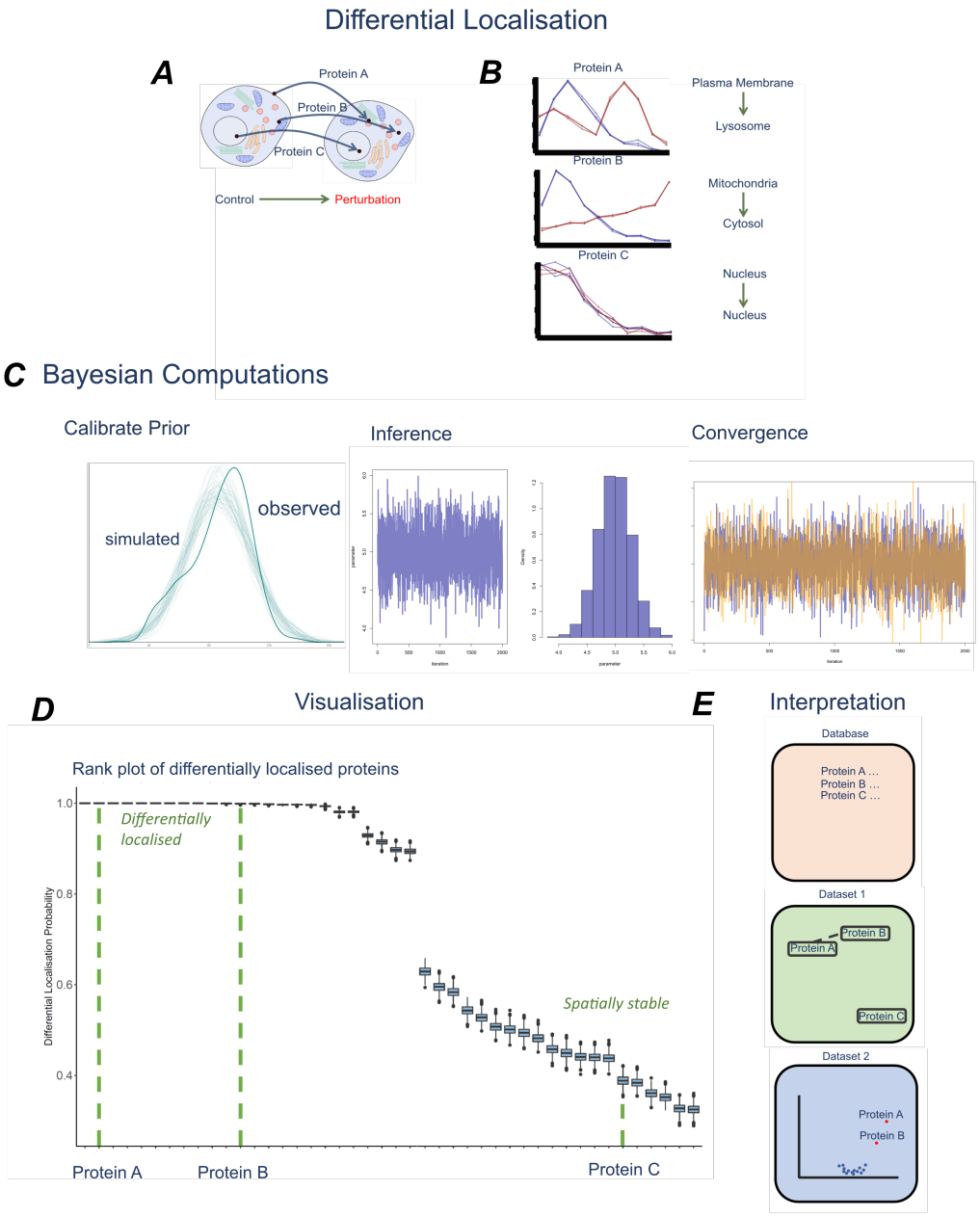

Figure 1: An overview of the BANDLE workflow. (A) A motivated differential localisation experiment is set-up with a perturbation of interest (B) Mass-spectrometry based spatial proteomics methods are applied to generate the data. (C) BANDLE is applied by first calibrating the prior, then performing inference using MCMC, as well as algorithmic assessing convergence. (D) The major results of BANDLE are represented in a rank plot. (E) Results are interpreted using auxiliary data or additional experiments.

2.2 Simulations demonstrate superior performance of BANDLE

2.2.1 BANDLE reduces false positives and increases power

To assess the performance of BANDLE and the MR approach, we run a number of simulations allowing us to ascertain the difference between each method in scenarios where we know the ground truth. We first start with a real dataset on Drosophila embryos and simulate replicates, as well as 20 protein re-localisations (Tan *et al.*, 2009). To simulate these datasets a bootstrapping approach is used, coupled with additional noise effects. The first simulation uses a simple bootstrapping approach, where a niche-specific noise component is included (see supplementary methods). The subsequent simulations start with the basic bootstrapping approach and add additional effects. The second and third simulations add batch effects: random and systematic respectively (see supplementary methods). Whilst the forth simulation generates misaligned features across datasets by permuting them (fraction swapping) - this models misaligned fractions between replicates (see supplementary methods). The final simulation includes both batch effects and feature permutations. The simulations are repeated 10 times, where each time we simulate entirely new datasets and re-localisations - this is repeated for each simulation task. We assess the methods on two metrics - the area under the curve (AUC) of the true positive rate and false positive rate for the detection of differential localised proteins. Furthermore, we determine the number of correctly differentially localised proteins at fixed thresholds (see supplementary methods).

Our proposed method, BANDLE, significantly outperforms the MR method with respect to AUC in all scenarios (t-test $p < 0.01$). Furthermore, it demonstrates that BANDLE is robust to a variety of situations, including batch effects. The performance of BANDLE based on the Dirichlet prior is already very good and thus it is unsurprising that we do not observe any significant improvements in AUC by including prior information on correlations captured by the Pólya-Gamma prior. Additional comparisons are made in the supplement where we make similar observations

The improved AUC, which demonstrates improved control of false positives and increased power, translates into increased discovery of differentially localised proteins. Indeed, BANDLE with the Dirichlet prior discovers around twice as many such re-localising proteins. Allowing prior correlations through the Pólya-Gamma prior demonstrates that additional differentially localised proteins are discovered. This is an important reality of those performing comparative and dynamic spatial proteomics experiments, since the experiments become more worthwhile with additional biological discoveries. In practice, the authors of the MR approach advocate additional replicates to calibrate which thresholds are used to declare a protein differentially localised. This assumes that the perturbation of interest does not have a strong effect on the properties of the sub-cellular niches, which restricts applicability. In contrast, BANDLE does not need additional mass-spectrometry experiments to calibrate its probabilistic ranking meaning more discoveries are made at lower cost.

In the following section, we examine the differences between the approaches in a simulated example. There we focus on the output, interpretation and statistical qualities of each approach, rather than the predictive performance of the methods.

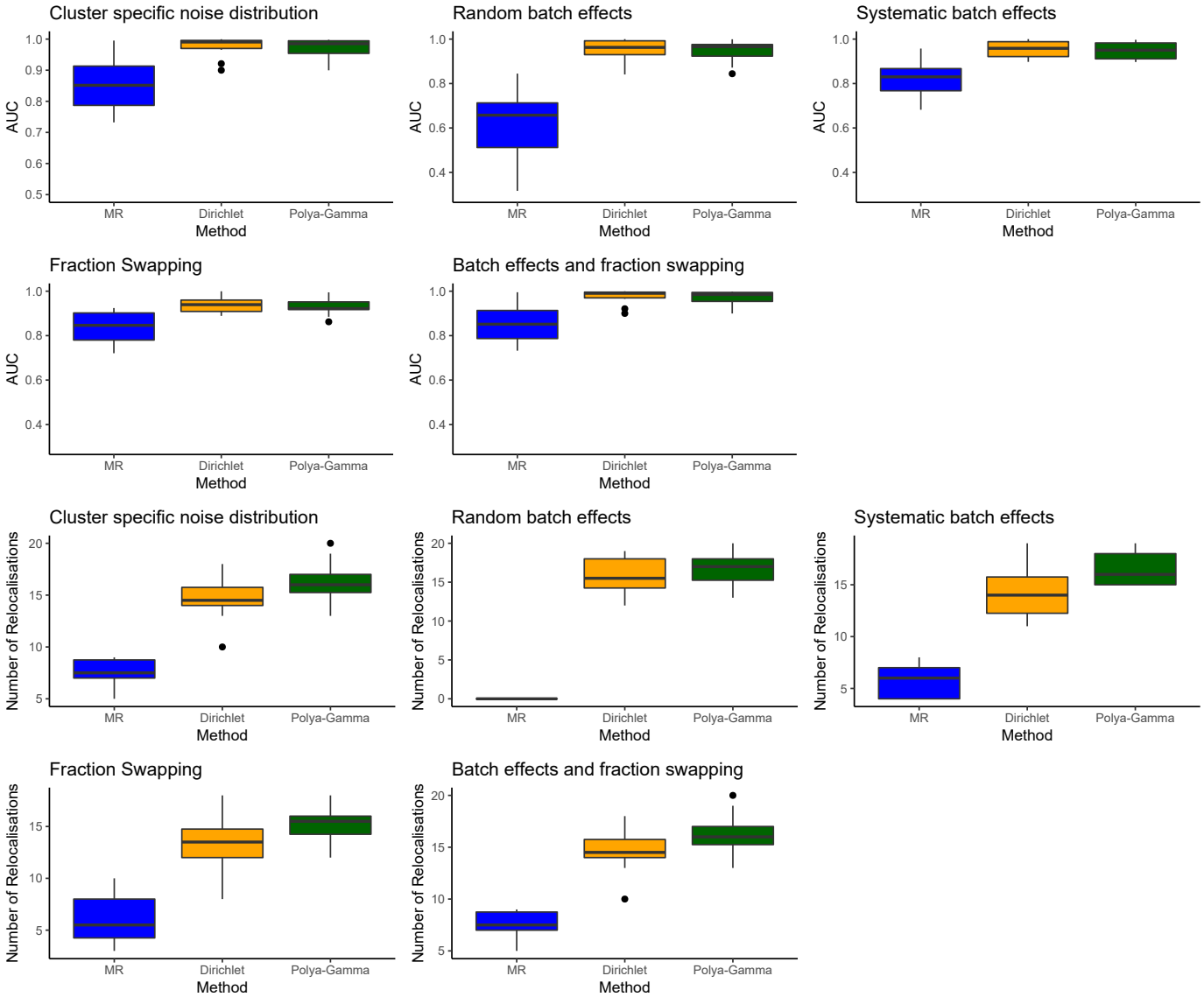

Figure 2: Boxplots comparing the performance of the MR approach and our proposed method BANDLE. BANDLE is separated into whether a Dirichlet-based prior was used or if the Polya-Gamma augmentation was applied. Each boxplot correspond to a different simulation scenario. The first 5 boxplots show BANDLE has significantly improved AUC in all scenarios. These AUCs are translated into the correct number of re-localisations and we can see that our method clearly outperforms the MR approach.

2.2.2 BANDLE quantifies uncertainty and is straightforward to interpret

In this section, we further explore the application of BANDLE with a Dirichlet prior and the MR approach, focusing on the interpretation and statistical properties of the two methods. Again, we simulate dynamic spatial proteomics data, starting from the Drosophila experiment in the scenario in which the MR method performed best. This is where there are cluster specific noise distributions but no other effects, such as batch effects, were included. Sample PCA plots of the data are presented in figure 3 A. There is a clear pattern of localisations across the data where proteins with know sub-cellular localisations are closer to each other. However, the organelle distributions clearly overlap and in some cases are highly dispersed - a representation of the challenges faced in real data. These data are annotated with 11 sub-cellular niches and 888 proteins are measured across 3 replicates of control and 3 of treatment (totalling 6 experiments). Re-localisations are simulated for 20 proteins.

We first apply the MR method according to the methods in Itzhak *et al.* (2016, 2017). We provide a brief description of the approach with full details in the methods. To begin, the difference profiles are computed by subtracting the quantitative values for each treatment from each control. Then the squared Mahalanobis distance is computed to the centre of the data and under a Gaussian assumption the null hypothesis is that these distance follow a Chi-squared distribution, ergo a p-value is obtained. This process is repeated across the 3 replicates and the largest p-value was then cubed and then corrected from multiple hypothesis testing using the Benjamini-Höchberg procedure (Benjamini and Hochberg, 1995). A negative $\log_{10}$ transform in then performed to obtain the M-score. To produce the R-score, Pearson correlations are computed between each difference profile for all pairwise combination of difference profiles. The lowest of the three R-scores is reported. The M-score and R-score are plotted against each other (see figure 3 B) and the proteins with high M-score and high R-score are considers "hits".

There are a number of assumptions underlying the MR methodology. Firstly, comparing difference profiles pairwise assumes that the features in both datasets exactly correspond. However, this precludes any stimuli that changes the biochemical properties of the organelles, since changing these properties may result in differing buoyant densities or pelleting of niches at different centrifugation speeds. Thus, whether density-gradient or differential centrifugation is used for organelle separation this assumption must be carefully assessed. Secondly, the Gaussian assumption ignores the natural clustering structure of the data because of the different organelle properties. Indeed, examination of the p-value distributions in a histogram (figure 3 C) shows that it clearly deviates from the mixture of distributions expected (p-values are uniformly distributed under the null). The peaking of p-value towards 1 suggests poor distributional assumptions (Holmes and Huber, 2018). Thus perhaps the Chi-squared distribution is a poor fit for the statistic of interested. Exploring this further, we fit a Chi-squared and Gamma distribution empirically to the statistics using maximum likelihood estimation (MLE) of the parameters. Figure 3 D show that the Gamma distribution is a better distributional fit - successfully capturing the tail behaviour of the statistic. The Chi-squared family is nested in the Gamma family of distributions, so if the theoretical Chi-squared distribution was a good fit the distributions would overlap. For a quantitative assessment of model fits we compute the negative log-likelihood of the data given the optimal distributions - the Gamma distribution has a markedly lower negative log-likelihood (Figure 3 E). This provides strong evidence that the underlying Gaussian assumptions are likely violated. Thirdly, it is inappropriate to cube p-values:

to combine p-value across experiments one could use Fisher's method (Mosteller and Fisher, 1948; Brown, 1975; Kost and McDermott, 2002) or the Harmonic mean p-value (HMP) (Good, 1958; Wilson, 2019) depending on the context. Indeed, the cube of the p-value is no longer a p-value. To elaborate, if $\mathcal{P}$ are a set of p-values, then under the assumption of the null hypothesis $\mathcal{P}$ is uniformly distributed; however, the cube is clearly not uniformly distributed. Since we no longer work with p-values, Benjamini-Höchberg correction becomes meaningless in this context. Transforming these values to a "Movement score", conflates significance with effect size which confounds data interpretation. Finally, summarising to a single pair of scores ignores their variability across experimental replicates.

BANDLE first models each sub-cellular niche non-parametrically (since the underlying functional forms are unknown (Crook *et al.*, 2019b)). Visualisation of the posterior predictive distributions from these fits for selected sub-cellular niches is given in figure 3 H - we observe a good correspondence between the model and the data. We can see that the different sub-cellular niches have contrasting correlation structures and thus niche specific distributions are required. These distributions are specific for each replicate of the experiment and also the two experimental conditions. The information from the replicates, and the control and treatment are combined using an integrative mixture model. Briefly, mixing proportions are defined across datasets allowing information to be shared between the control and treatment (see methods for more details). This formulation allows us to compute the probability that a protein is assigned to a different sub-cellular niche between the two experiments - the *differential localisation probability*. The proteins can then be ranked from most probably differentially localised to least (figure 3 H). The figure is simple to interpret: the proteins with highest rank are the most likely to have differentially localised between the experiment, having been confidently assigned to different sub-cellular niches in the control versus treatment. The proteins with lowest rank are highly unlikely to have moved during the experiment - the localisation are stable. This is important information in itself, especially when combined with other information such as changes in abundance or post translational modification. Figure 3 G shows the 30 proteins with highest rank visualising the uncertainty in the differential localisation probability (see methods). This ranking allows us to prioritise which proteins to follow up in validation experiments. The ranking can also be mapped onto other experimental data, such as expression or protein-protein interaction data. The probabilistic ranking produced by BANDLE is more closely aligned with the phenomenon of interest. Indeed, we divide the data into the proteins that were differentially localised and those that were not. Then from plotting the distribution of the statistics from the respective methods, it is clear that output from BANDLE is most closely associated with re-localisation events (figure 3 I).

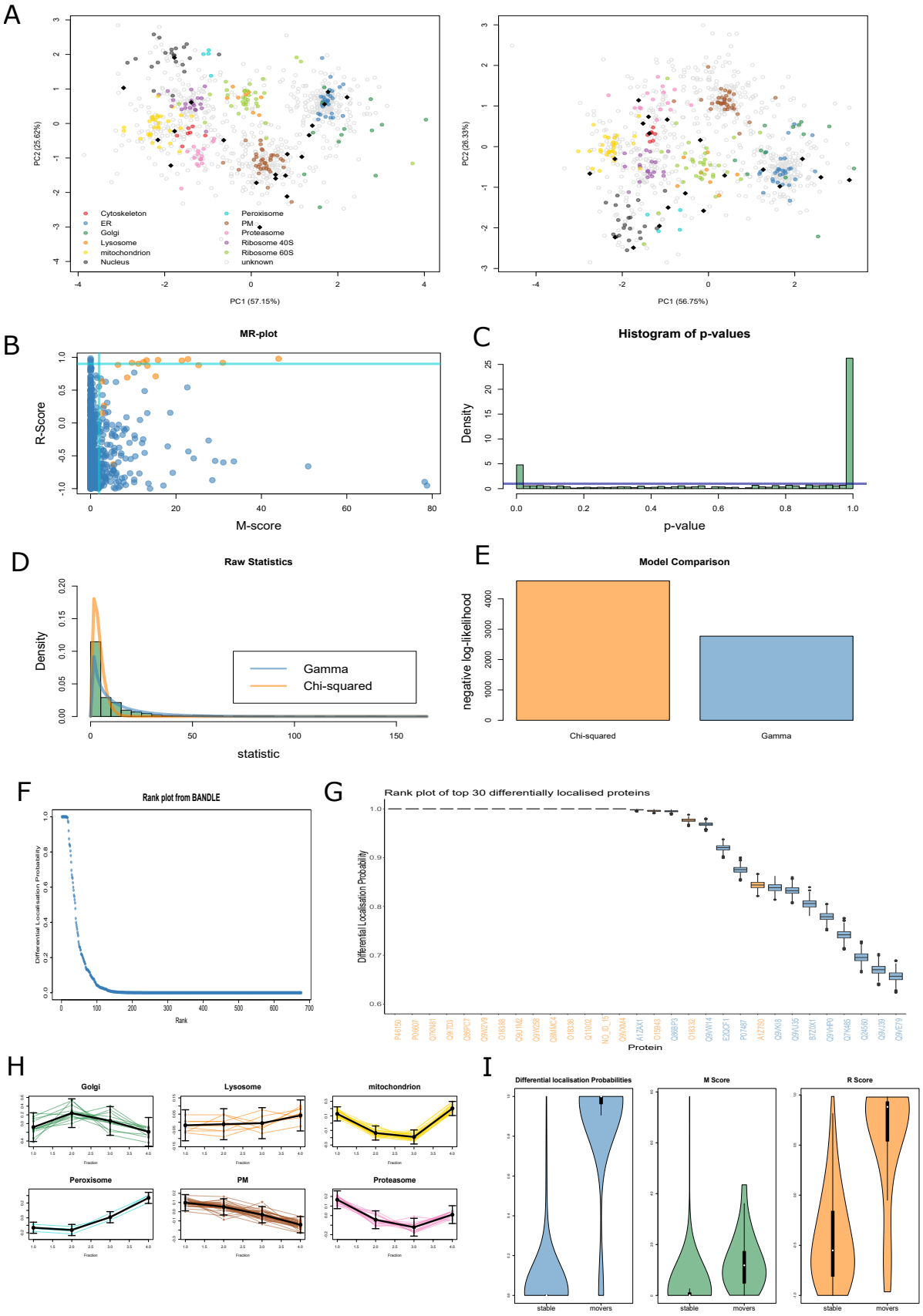
A
PC2 (25.62%)
PC1 (57.15%)
Cytoskeleton
ER
Golgi
Lysosome
mitochondrion
Nucleus
Peroxisome
PM
Proteasome
Ribosome 40S
Ribosome 60S
unknown
PC2 (26.33%)
PC1 (56.75%)
B
MR-plot
R-Score
M-score
C
Histogram of p-values
Density
p-value
D
Raw Statistics
Density
statistic
Gamma
Chi-squared
E
Model Comparison
negative log-likelihood
Chi-squared
Gamma
F
Rank plot from BANDLE
Differential Localisation Probability
Rank
G
Rank plot of top 30 differentially localised proteins
Differential Localisation Probability
Protein
H
Golgi
Lysosome
mitochondrion
Peroxisome
PM
Proteasome
Fraction
I
Differential localisation Probabilities
M Score
R Score
stable
movers

Figure 3:

Figure 3: (A) Example PCA plots where pointers correspond to proteins. Marker proteins are coloured according to their subcellular niche, whilst proteins with unknown localisation are in grey. Simulated translocations are highlighted in black, where the left corresponds to control and right to the perturbed dataset. (B) An MR-plot showing movement score against reproducibility score. Each pointer correspond to a protein and orange pointers correspond to simulated translocations and blue otherwise. Teal lines are drawn at suggested thresholds with proteins in the top right corner considered hits. (C) A p-value histogram from the statistic underlying the MR-method. A purple line indicates uniformity. This histogram clearly deviates from uniform behaviour. (D) A histogram of the raw statistics underlying the MR method. A Chi-square (orange) and Gamma (blue) fit are overlaid (obtained using maximum likelihood estimation). The Gamma distribution clearly captures the tail behaviour. (E) model selection on the raw statistic using the Chi-squared and Gamma models. The Gamma model has lower negative log-likelihood and is thus a better model fit. (F) A BANDLE rank plot where proteins are ranked from most to least likely to differentially localised. The differentially localisation probability is recorded on the y-axis. (G) A BANDLE rank plot of the top 30 differentially localised proteins with uncertainty estimates for the differential localisation probability. Proteins marked in orange were simulated translocations. (H) Posterior predictive distributions (black) overlayed on the marker profiles for different subcellular niches showing the quality of the non-parametric BANDLE fits. (I) Violin plots for the differential localisation probabilities, the M score and R score. The distribution are split between differential localised (movers) and spatially stable proteins. Clearly, the differential localisation probabilities correlate most closely with the phenomena of interest.

2.3 Applications to differential localisation experiments

2.3.1 Characterising differential localisation upon EGF Stimulation

Having carefully assessed the statistical properties of our approach, BANDLE, and the MR method, we apply these approaches to a number of datasets. First, we consider the *Dynamic Organeller Maps* (DOMs) dataset of Itzhak *et al.* (2016), exploring the effects of EGF stimulation in HeLa cells. In this experiment, SILAC labelled HeLa cell were cultured and recombinant EGF was added to the culture at a concentration of 20 ng ml^{-1} (see (Itzhak *et al.*, 2016)). A total of 2237 complete protein profiles were measured across 3 replicates of control and 3 replicates of EGF treated HeLa cells. Principal Component Analysis (PCA) projections of the data can be visualised in the supplement. A quality control assessment was performed using the approach of Gatto *et al.* (2019). As a result, Nuclear pore complex, peroxisome and Golgi annotations were removed, since the marker proteins of these classes were highly dispersed.

The MR method was applied as described in the methods and the results can be visualised in figure 4 C. 7 proteins are predicted to be differentially localised with the MR method with the thresholds suggested by (Itzhak *et al.*, 2016). These include 3 core proteins of the EGF signalling pathway SHC1, GRB2 and EGFR (Oda *et al.*, 2005) and other, potentially related, proteins TMEM214, ACOT2, AHNAK, PKN2. Since the MR approach does not provide information about how the functional residency of the proteins change, it is challenging to interpret these results without further analytical approaches.

To quantify uncertainty and gain deeper insight into the perturbation of HeLa cell after EGF stimulation we applied our BANDLE pipeline. Firstly, the rank plots displays a characteristic shape suggesting that most proteins are unlikely to be differential localisation upon EGF stimulation (figure 4 D). Furthermore, we provide uncertainty estimates in the probability that a protein is differentially localised for select top proteins (figure 4 E). Furthermore, we visualise the change in localisation for the proteins known to re-localise upon EGF stimulation: SHC1, GRB2 and EGFR (figure 4 G). This is displayed by projecting the posterior localisation probabilities on to the corresponding PCA coordinates. These probabilities are then smoothed using a Nadaraya-Watson kernel estimator (Nadaraya, 1964; Watson, 1964) and visualised as contours.

Given the well-documented interplay between phosphorylation and sub-cellular localisation (Lee *et al.*, 2012; Christian *et al.*, 2016; Puertollano *et al.*, 2018; Balta *et al.*, 2018), we hypothesised that proteins with the greatest differential phosphorylation would correlate with proteins that were more likely to be differentially localised. To this end, we integrated our analysis with a time-resolved phosphoproteomic dataset of EGF stimulation using MS-based quantitation (Köksal *et al.*, 2018). In their study, EGF stimulated cells were cultured to 8 different time points: $0, 2, 4, 8, 16, 32, 64, 128$ mins. For MS-based quantitation trypsin digested peptides are laballed using iTRAQ 8-plex and pooled. Immunoprecipitation was used to enrich for phosphorylated tyrosine residues (Possemato *et al.*, 2017) and the enrichment of phosphosites on serine and threonine residues was performed via immobilized metal affinity chromatography (IMAC) (Ficarro *et al.*, 2002; Moser and White, 2006).

For each phosphopeptide corresponding to a unique protein, we computed the largest log_2 fold change observed across the time course. Given that the changes in localisation occur within 20 minutes, we restricted ourselves to the first 6 time points (Itzhak *et al.*, 2016). We then took the top 10 proteins ranked by each of the MR method and BANDLE. These rankings are then correlated

with rankings obtained from the changes in phosphorylation. The Spearman rank correlations were recomputed for $5,000$ bootstrap resamples to obtain bootstrap distributions of correlations (see figure 4). We report the mean correlation and the 95% boostrap confidence intervals. The correlation between the ranks of the MR method and the phosphoproteomic dataset was $\rho_S = 0.40\,(-0.49, 0.85)$, whilst the the correlation when using the ranking of BANDLE was $\rho_S = 0.68\,(0.02, 0.98)$. Alongside the statistical and interpretable benefits of BANDLE, it is clear the approach has the utility to provide insight into localisation dynamics.

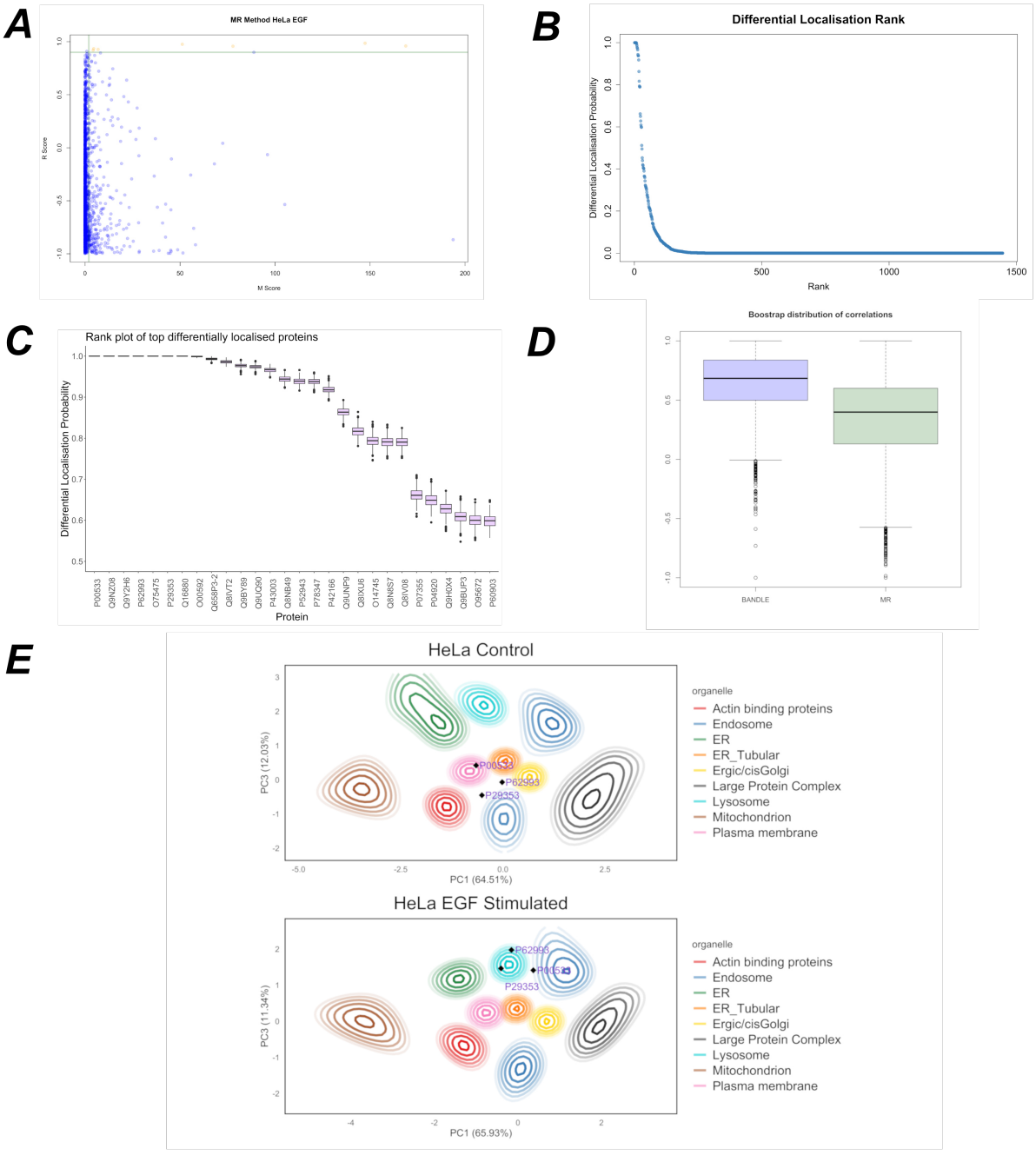
A
MR Method HeLa EGF
R Score
M Score
B
Differential Localisation Rank
Differential Localisation Probability
Rank
C
Rank plot of top differentially localised proteins
Differential Localisation Probability
Protein
P00533
Q8NZ08
Q8Y2H6
P62993
O75475
P29353
Q16880
O00592
Q658P3-2
Q8VT2
Q9BY89
Q9UQ90
P43003
Q8NB49
P52943
P78347
P42166
Q9UNP9
Q8IXU6
O14745
Q8N8S7
Q8IV08
P07355
P04920
Q9H0X4
Q9BUP3
O95672
P60903
D
Boostrap distribution of correlations
BANDLE
MR
E
HeLa Control
PC3 (12.03%)
PC1 (64.51%)
P00533
P62993
P29353
organelle
Actin binding proteins
Endosome
ER
ER_Tubular
Ergic/cisGolgi
Large Protein Complex
Lysosome
Mitochondrion
Plasma membrane
HeLa EGF Stimulated
PC3 (11.34%)
PC1 (65.93%)
P62993
P29353
organelle
Actin binding proteins
Endosome
ER
ER_Tubular
Ergic/cisGolgi
Large Protein Complex
Lysosome
Mitochondrion
Plasma membrane

Figure 4:

Figure 4: (A) An MR-plot where dark green lines are drawn at suggested threshold and hits are highlighted in orange. (B) BANDLE rank plot showing the distribution of differentially localised proteins. (C) The top differentially localised proteins from BANDLE plotted with uncertainty estimates. (D) Boostrap distributions of correlations with a phosphoprotemomic time-course experiment. The BANDLE confidence intervals differ significantly from 0, whilst the MR method do not. (E) PCA plots with (smoothed) localisation probabilities project onto them. Each colour represent an organelle and ellipses represent lines of isoprobability. The inner ellipse corresponds to 0.99 and the proceed line 0.95 with further lines decreasing by 0.05 each time. The protein are highlight demonstrating example relocalisations. EGFR (P005330) clearly relocalises from the PM to endosome, whilst SHC-1 (P29353) and GRB2 (P62993) relocalise from unknown localisation to the Lysosome.

2.3.2 BANDLE obtains deeper insights into AP-4 dependent localisation

The adaptor protein (AP) complexes are a set of heterotetrameric complexes, which transport transmembrane cargo protein vesicles (Robinson, 2015). The AP1-3 complexes are well characterised: AP-1 mediates the transport of lysosomal hydrolases from the trans-Golgi to the endsomes (Karin *et al.*, 1997; Hess *et al.*, 2004); AP-2 has a significant role in the regulation of endocytosis (Motley *et al.*, 2003); AP-3 is involved in the sorting of trans-Golgi proteins targeted to the lysosome (Dell'Angelica *et al.*, 1998). The role of the AP-4 complex has become better understood in recent years (Hirst *et al.*, 1999, 2013; Mattera *et al.*, 2017; Davies *et al.*, 2018; De Pace *et al.*, 2018; Behne *et al.*, 2020; Ebrahimi-Fakhari *et al.*, 2020; Ivankovic *et al.*, 2020) and is of noted interest because loss-of-function mutations resulting in early-onset progressive spastic paraplegia (Moreno-De-Luca *et al.*, 2011). The altered sub-cellular distribution of ATG9A, as a result of loss-of-functions AP-4 mutation, (Mattera *et al.*, 2017; Davies *et al.*, 2018; De Pace *et al.*, 2018) is believed to be a key contributor to the pathology of AP-4 deficiency syndrome (Mattera *et al.*, 2017; Davies *et al.*, 2018; De Pace *et al.*, 2018; Behne *et al.*, 2020; Ebrahimi-Fakhari *et al.*, 2020; Ivankovic *et al.*, 2020).

AP-4 consists of four subunits ($\beta 4$, ε, $\mu 4$ and $\sigma 4$) forming an obligate complex (Dell'Angelica *et al.*, 1998). Davies *et al.* (2018) study the functional role of AP-4 using spatial proteomics; in particular, the DOM workflow mentioned previously. As part of their study, they use AP-4 CRISPR knockout cells to interrogate the effect on the spatial proteome when AP-4 function has been ablated.

The DOM experiment we re-analyse from Davies *et al.* (2018) provides full quantitative measurements for 3926 proteins across two replicates of wild-type cells and two replicates where the $\beta 4$ subunit has been knocked-out. They also produce two replicates where AP4E1 is knocked-out but this is not considered here for brevity. The data are visualise as PCA plots (see supplement). As in the previous analysis, we run a quality control step removing the Actin binding protein and Nuclear pore complex annotations (Gatto *et al.*, 2019). This dataset is particular challenging to analyse because there are only two replicates for control and treatment. The value of Bayesian analysis is the ability to provide prior information to regularise, as well as the quantification of uncertainty which is more critical in data sparse scenarios.

Previous application of the MR methods led to authors to find that SERINC 1 (Q9NRX5), SERINC 3 (Q13530) were differentially localised, as well as an altered subcellular distribution for ATG9A (Q7Z3C6) (Davies *et al.*, 2018). Their results suggest these are cargo proteins of the AP-4 complex that are packaged into vesicles at the trans-Golgi before being transported to the cell periphery. All together their results suggest AP-4 provides spatial regulation of autophagy and that AP-4 neurological pathology is linked to disturbances in membrane trafficking in neurons (Mattera *et al.*, 2017; Davies *et al.*, 2018).

We apply our method BANDLE in order to gain further insights into AP-4 dependent localisation. We compute the differential localisation probability; the associated uncertainty estimates and rank proteins according to this statistic (see figures 5 A and B). Characteristic S shape plots are observed with most proteins not differential localisation upon knock-out of AP-4 $\beta 4$. The results of both SERINC 1 and 3 are validated, as we compute a differential localisation probability greater than 0.95 for these proteins. Furthermore, 16 of the top 20 proteins are membrane-bound or membrane-associated proteins (FDR < 0.01 hyper-geometric test). To demonstrate the benefit of our probabilistic ranking, we perform two-sided KS rank test against the functional annotations provided in the STRING database (corrected for multiple testing within each functional framework)

(Szklarczyk *et al.*, 2019). We find that processes such as ER to Golgi transport and lipid metabolism are more highly ranked that would be expected at random (FDR < 0.01), as well as endosome and Golgi localisations (FDR < 0.01). Whilst processes associated with translation, ribosome localisation and function appear significantly lower in the ranking (FDR < 0.01). As expected, this provides a high level overview and evidence for the functional nature of AP-4 in the secretary pathway.

Taking a more precise view on our results, we examine the top 20 differentially localised proteins in more detail. We compute the Spearman correlation matrix between these proteins and observe clustering, suggesting the proteins act in a coordinated way (see figure 5 C). Visualising the data in a heatmap (figure 5 D), after mean and variance normalisation, we observe a highly concordant pattern: most proteins are enriched in fractions 4 and 5. These fractions are obtained from the highest centrifugation speeds and so differentially pellet light membrane organelles, such as endosomes and lysosomes (Itzhak *et al.*, 2016; Geladaki *et al.*, 2019). Again, further evidence for the role of AP-4 dependent localisation dynamics within the secretary pathway.

In figure 5 C, we observe a large cluster of 9 proteins, which includes SERINC 1 and 3. Amongst these 9 proteins is SLC38A2, a ubiquitously expressed amino-acid transporter that is widely express in the central nervous system and is recruited to the plasma membrane from a pool localised in the trans-Golgi (Hatanaka *et al.*, 2000; Bevilacqua *et al.*, 2005; Gonzalez-Gonzalez *et al.*, 2005; Melone *et al.*, 2006). Thus, its suggested differential localisation here provides further evidence for the role of AP-4 as a membrane trafficker from the trans-Golgi. Another protein in this cluster is TMEM 199 (Q8N511) a protein of unknown function that is involved in lysosomal degradation (Miles *et al.*, 2017). Furthermore, it has been implicated in Golgi homoeostasis but the functional nature of this process is unknown (Jansen *et al.*, 2016). Probing further, we observe that TMEM199 acts in a coordinated fashion with SERINC 1 and 3. Marked re-localisations are observed on PCA plots toward the endo/lysosomal regions (see figure 5 E) and we note that the quantitative profiles of SERINC 1, SERINC 3 and TMEM199 act in an analogous way upon AP-4 knockout (see figure 5 F). Our findings motivate additional studies to elucidate AP-4 dependent localisation and separate biological hits from potential clonal artefacts.

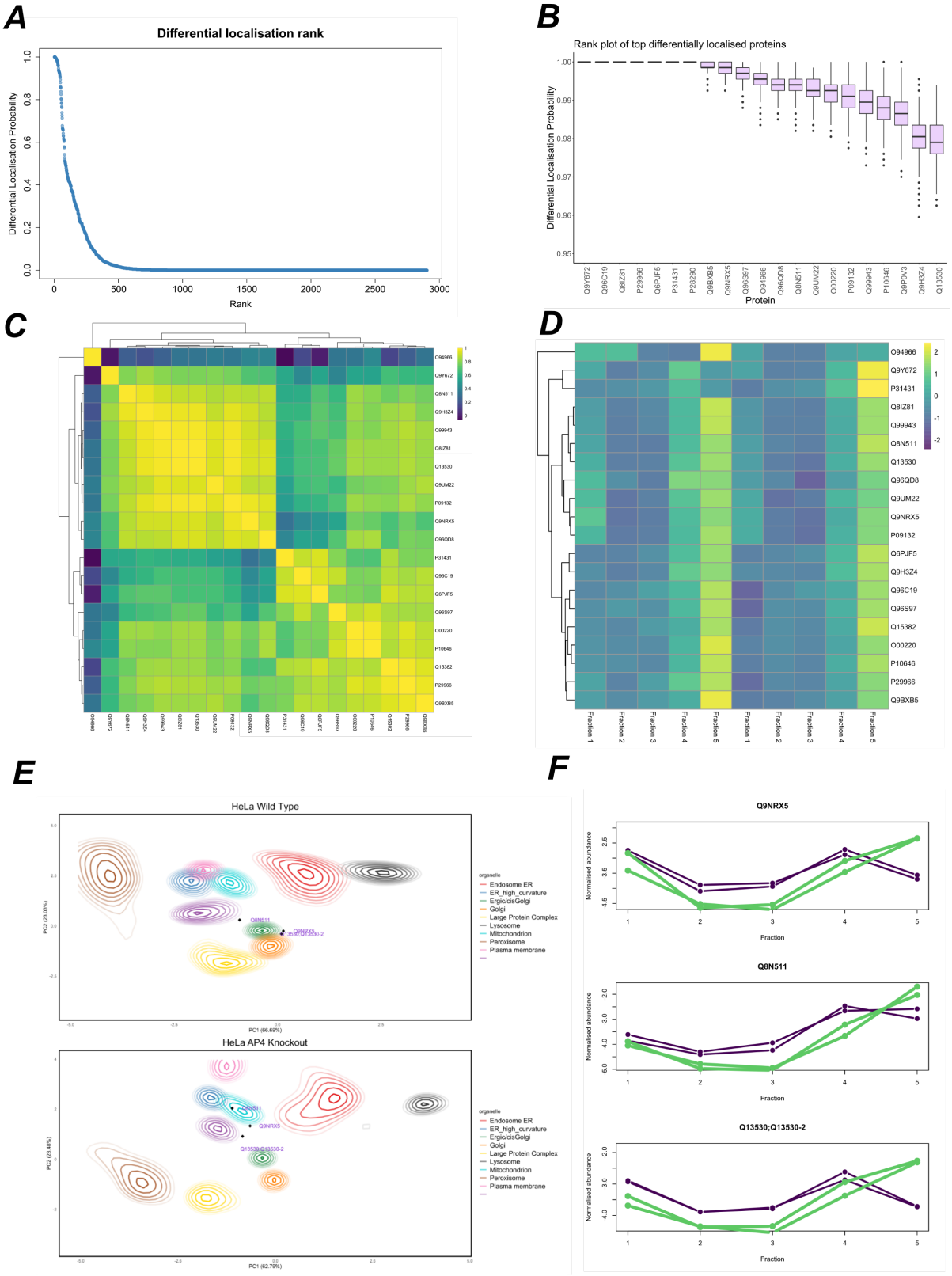
A
Differential localisation rank
Differential Localisation Probability
Rank
B
Rank plot of top differentially localised proteins
Differential Localisation Probability
Protein
C
D
E
HeLa Wild Type
HeLa AP4 Knockout
organelle
Endosome ER
ER_high_curvature
Ergic/cisGolgi
Golgi
Large Protein Complex
Lysosome
Mitochondrion
Peroxisome
Plasma membrane
F
Q9NRX5
Q8N511
Q13530;Q13530-2
Normalised abundance
Fraction

Figure 5:

Figure 5: (A) BANDLE rank plot showing the distribution of differentially localised proteins. (B) The top differentially localised proteins from BANDLE plotted with uncertainty estimates. (C) A Spearman correlation heatmap showing strong correlations and coclustering behaviour of proteins that have AP-4 dependent localisation (D) Normalised mass-spectrometry profiles plotted as a heatmap from the AP-4 knockout data. Proteins are shown to have similar behaviour with greater intensity in fraction 5, where light membrane organelles are likely to pellet. (E) PCA plots with (smoothed) localisation probabilities project onto them. Each colour represent an organelle and ellipses represent lines of isoprobability. The inner ellipse corresponds to 0.99 and the proceed line 0.95 with further lines decreasing by 0.05 each time. The proteins SERINC 1 and 3, as well as TMEM199 are highlight demonstrating example relocalisations. (F) Normalised abundance profiles showing that SERINC 1, SERINC 3 and TMEM199 show similar behaviour upon knockout of AP-4.

2.4 Rewiring the proteome under Cytomegalovirus infection

2.4.1 The host spatial-temporal proteome

Human Cytomegalovirus (HCMV) infection is a ubiquitous herpesvirus that burdens the majority of the populous (Cannon *et al.*, 2010). In healthy immune systems, HCMV establishes latent infection following initial viral communication (Reeves *et al.*, 2005) and reactivation can lead to serious pathology in certain imunno-compromised individuals (Boeckh and Nichols, 2004). HCMV has a highly expanded genome with vast capabilities to encode functional proteins (Murphy *et al.*, 2003; Stern-Ginossar *et al.*, 2012). For the virus to succeed it carefully modulates cellular functions *en masse* (Jean Beltran and Cristea, 2014).

Initial viral infection involves endocytosis of the viron into the cell (Isaacson *et al.*, 2008), host machinery is then used to transport viral capsids into the nucleus (Ogawa-Goto *et al.*, 2003). Within the host nucleus viral transcription and genome replication occurs (Milbradt *et al.*, 2007; Gibson, 2008; Kalejta, 2008). Meanwhile, other viral proteins are targeted to the secretory pathway to inhibit the host immune response and regulate the expression of viral genes (Stamminger *et al.*, 2002; Feng *et al.*, 2006; Hwang and Kalejta, 2007; Mitchell *et al.*, 2009; Cristea *et al.*, 2010; Li *et al.*, 2013), rewire signalling pathways (Yurochko, 2008) and modulate metabolism (Yu *et al.*, 2011). In later phases, the cellular trafficking pathways and the secretory organelles are hijacked for the formation of the viral assembly complex (vAC) (Buchkovich *et al.*, 2010; Moorman *et al.*, 2010; Alwine, 2012; Das *et al.*, 2007; Das and Pellett, 2011). Due to the diversity of cellular processes manipulated during HCMV infection, it is often used as a paradigm to analyse virus-host interactions (Weekes *et al.*, 2014).

There has been a recent flurry in applying system-wide proteomic approaches to the HCMV infection model. Weekes *et al.* (2014) developed *quantitative temporal viromics* a multiplexed proteomic approach to understand the temporal response of thousands of cellular host and viral proteins. More recently, to discover proteins involved in the innate immune response, a multiplexed proteasome-lysosome degradation assay found that more than 100 host proteins are degraded shortly after viral-infection (Nightingale *et al.*, 2018). Meanwhile, a comprehensive mass spectrometry interactome analysis has identified thousands of host-virus interactions (Nobre *et al.*, 2019). Furthermore, high-throughput temporal proteomic analysis has revealed the importance of protein acetylation (post-translational modification of lysine amino acids), as an integral component during HCMV infection (Murray *et al.*, 2018).

Beltran *et al.* (2016) use spatial and temporal proteomics to investigate the response of the human host proteome to HCMV infection. The authors perform subcellular fractionation on uninfected (control) and HCMV infected (treated) cells at 5 different time point $(24, 48, 72, 96, 120)$ hours post infection (hpi). The authors then used neural networks to classify proteins to sub-cellular niches at each time point in the control and treated cells, allowing a descriptive initial analysis of the data. Proteins with differential classification at each time point are those that are believed to be differential localised. However, the challenge of this study is that only a single replicate is produce in each situation. This renders the MR method of (Itzhak *et al.*, 2016) inapplicable.

Differential classification is a reasonable approach to probe differential localisation though it neglects information shared across both experiments and it is not quantitative (i.e. no p-value or posterior probability of change). In the case of single replicates, by sharing information and pro-

viding prior information we are able to improve inference and obtain deeper insights. We apply BANDLE to control and HCMV-treated cells at 24 hpi, in the interest of brevity, to explore further the host spatial-temporal proteome. Our analysis reflects extensive rewiring of the proteome with hundreds of proteins differentially localised on HCMV infection. We highlight an example of differential localisation with SCARB1 (see figure 6 A), with a localisation in the secretory pathway shifting toward a PM/cytosolic localisation, similar to what has previously be observed (Beltran *et al.*, 2016).

To obtain global insights into the functional behaviour of the differentially localised proteins, we performed a Gene Ontology (GO) enrichment analysis. An extensive list of terms is enriched and these can be divided broadly into subcategories such as translation and transcription; transport; viral processes; and immune process (see supplement). These results reflect closely the early phase of HCMV infection (Jean Beltran and Cristea, 2014). Pathway enrichment analysis highlights terms related to a viral infection (Viral mRNA Translation, Influenza Life Cycle, Infectious disease, Innate Immune System, Immune System, MHC class II antigen presentation, Antigen processing-Cross presentation, Host Interactions of HIV factors, HIV Infection) (see figure 6 B). Pathway analysis also reveals known processes that are modulated during HCMV infection, such as membrane trafficking (Bozidis *et al.*, 2010; Niemann *et al.*, 2014; Zeltzer *et al.*, 2018), Extracellular matrix organization (Reinhardt *et al.*, 2006) and rab regulation of trafficking (Lučin *et al.*, 2018).

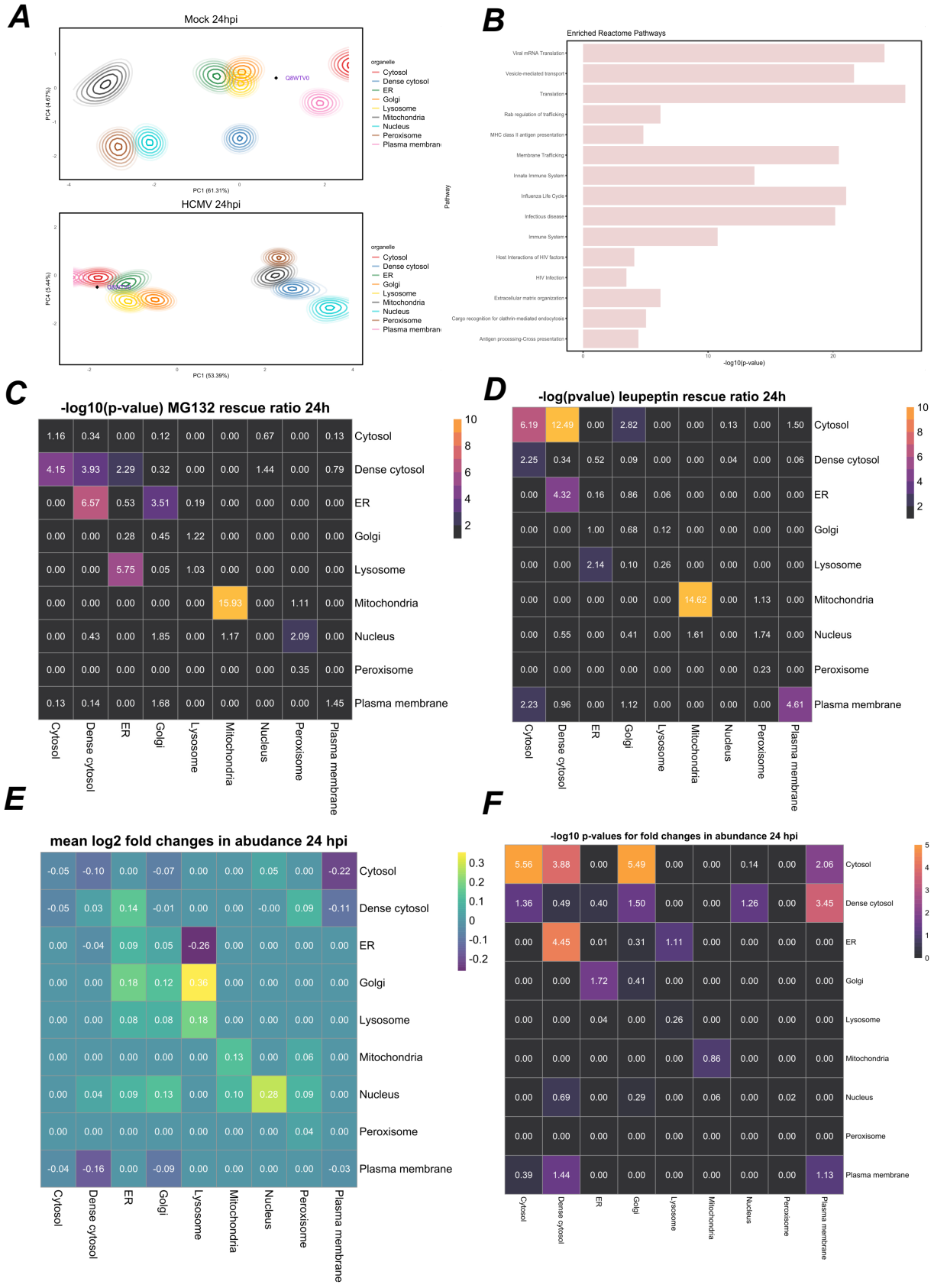
A
Mock 24hpi
Q8WTV0
PC1 (61.31%)
PC4 (4.67%)
organelle
Cytosol
Dense cytosol
ER
Golgi
Lysosome
Mitochondria
Nucleus
Peroxisome
Plasma membran
HCMV 24hpi
PC1 (53.39%)
PC4 (5.44%)
B
Enriched Reactome Pathways
Viral mRNA Translation
Vesicle-mediated transport
Translation
Rab regulation of trafficking
MHC class II antigen presentation
Membrane Trafficking
Innate Immune System
Influenza Life Cycle
Infectious disease
Immune System
Host Interactions of HIV factors
HIV Infection
Extracellular matrix organization
Cargo recognition for clathrin-mediated endocytosis
Antigen processing-Cross presentation
Pathway
-log10(p-value)
C
-log10(p-value) MG132 rescue ratio 24h
D
-log(pvalue) leupeptin rescue ratio 24h
E
mean log2 fold changes in abudance 24 hpi
F
-log10 p-values for fold changes in abundance 24 hpi
Cytosol
Dense cytosol
ER
Golgi
Lysosome
Mitochondria
Nucleus
Peroxisome
Plasma membrane

Figure 6:

Figure 6: (A) PCA plots with (smoothed) localisation probabilities project onto them. Each colour represent an organelle and ellipses represent lines of isoprobability. The inner ellipse corresponds to 0.99 and the proceed line 0.95 with further lines decreasing by 0.05 each time. The relocalisation of SCARB1 is highlighted on the plot (B) Reactome pathway enrichment results. (C) A heatmap representation of the MG132 inhibitor degradation data at 24 hpi. $\log_{10} p$-values are overlaid onto the spatial patterns across MOCK and HCMV infected cell 24hpi. The y-axis corresponds to localisation in the MOCK dataset whilst the x-axis corresponds to HCMV infected cells. (D) As for C but for the leupeptin inhibitor. (E) Mean log2 abundance fold changes are overlaid on a heatmap according to their spatial pattern (F) The p-values corresponding to the fold changes observed in E.

2.4.2 Integrating HCMV proteomic datasets to add functional relevance to spatial proteomics data

The spatial information obtained here allows us to perform careful integration with other high-resolution proteomic datasets. The degradation screens by Nightingale *et al.* (2018) identified proteins that were actively degraded during HCMV infection but gave no information regarding the spatial location of the targets. To determine the location of host proteins targeted by HCMV for degradation, the BANDLE revised spatial data at 24 hpi was overlapped with proteins that were degraded by the proteasome or lysosome. The subcellular location of the host proteins is displayed for the 24 h timepoint. To determine the spatial granularity of the degradation data we tested whether the proteins assigned to each spatial pattern had a significantly different degradation distribution that the distribution of all proteins in the experiment (t-test). We note that proteins that are differentially localised are no more likely to be targeted for degradation and those that are not (see supplement).

Degradation data from Nightingale *et al.* (2018) are overlaid as a heatmap, showing a $-\log_{10}(p\text{-value})$ for each inhibitor (figure 6 C and D). For proteasomal targeted proteins (MG132), the data highlight a high number of proteins degraded from the mitochondria. The mitochondria act as a signalling platform for apoptosis and innate immunity and it is already well established that HCMV can subvert these processes to its advantage (Crow *et al.*, 2016). Furthermore, there is a high degree of protein degradation as one might expect in proteasome fractions (dense cytosol), with an enrichment of proteins recruited from the ER and cytosol (see supplement). For lysosomal targeted proteins (leupeptin) there was a high degree of proteins degraded from the mitochondria, cytosol and plasma membrane. There were also several proteins degraded that moved from the cytosol to the dense cytosol.

Many host proteins are up or down regulated upon HCMV infection (Weekes *et al.*, 2014). We examine more recent abundance data from Murray *et al.* (2018) at 24 hpi and first we note that differentially localised proteins are not more abundant than spatially stable proteins (see supplement). However, we see a strong spatial pattern when we overlay the abundance pattern on a heatmap. In figure 6 E , we report the mean log2 fold change for proteins stratified according to predicted subcellular localisation. It is important to combine spatial and abundance data, since a differentially localised protein may not undergo a true translocation event but rather a new pool of proteins is synthesised. The significance of these abundance changes is highlighted in figure 6 F. For example, there is a significant decrease in the abundance of the protein recruited to the dense cytosol from the ER (see supplement). Some of the larger changes are not significant because there are too few proteins with the same spatial pattern. We note that FAM3C, a protein involved in platelet degranulation, is upregulated at 24 hpi. Furthermore, FAM3C relocalises from the Golgi to the Lysosome, its Golgi localisation is in concordance with the Human Protein Atlas (HPA) (Thul *et al.*, 2017) and its Lysosome relocalisation suggests that it is trafficked through the secretory pathway before undergoing degranulation.

Upon integration of the acetylation data of Murray *et al.* (2018), the spatial patterns are much more nuanced (see figures 7 A and B). Perhaps surprisingly, we do not observe increased acetylation levels amongst differentially localised proteins (see supplement). The only significant pattern is for proteins relocalising from the dense cytosol to the cytosol; however, we observe this is driven by a single protein Skp1 (see supplement), which shows a 2.5-fold increase in acetylation at 24 hpi

for Skp1 and there is an increase in its RNA transcript at 24 hpi (Nightingale *et al.*, 2018). The Skp1 protein is part of an E3 ubiquitin ligase complex that targets proteins for degradation. E3 ligases are often manipulated by viruses in order to control cellular processes to create a cell states that benefit viral replication and survival (Mahon *et al.*, 2014). It is therefore possible that HCMV is controlling Skp1 activity through acetylation at its C-terminus, leading to its translocation and likely change in function.

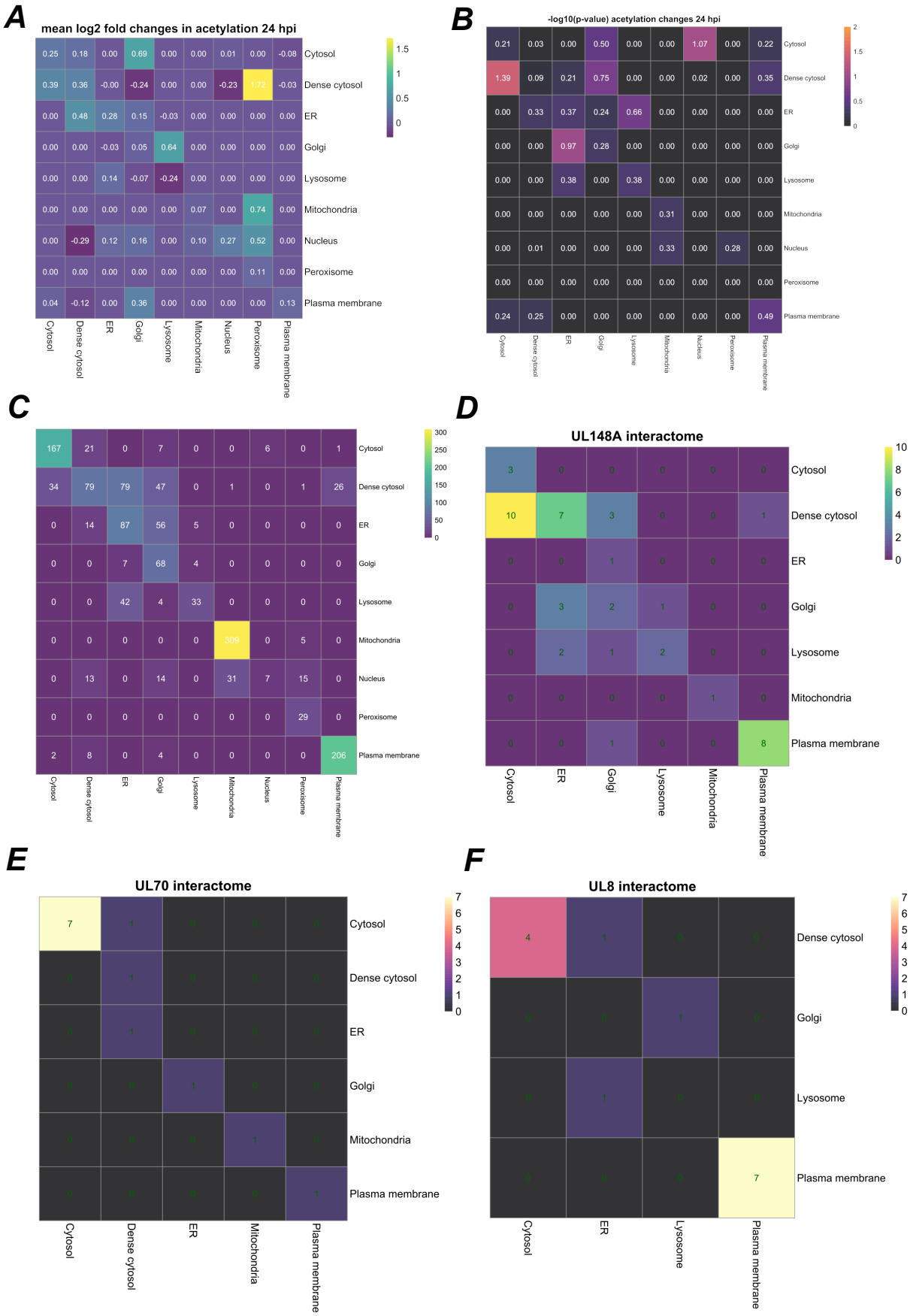
A
mean log2 fold changes in acetylation 24 hpi
Cytosol 0.25 0.18 0.00 0.69 0.00 0.00 0.01 0.00 -0.08
Dense cytosol 0.39 0.36 -0.00 -0.24 0.00 0.00 -0.23 1.72 -0.03
ER 0.00 0.48 0.28 0.15 -0.03 0.00 0.00 0.00 0.00
Golgi 0.00 0.00 -0.03 0.05 0.64 0.00 0.00 0.00 0.00
Lysosome 0.00 0.00 0.14 -0.07 -0.24 0.00 0.00 0.00 0.00
Mitochondria 0.00 0.00 0.00 0.00 0.00 0.07 0.00 0.74 0.00
Nucleus 0.00 -0.29 0.12 0.16 0.00 0.10 0.27 0.52 0.00
Peroxisome 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.11 0.00
Plasma membrane 0.04 -0.12 0.00 0.36 0.00 0.00 0.00 0.00 0.13
Cytosol Dense cytosol ER Golgi Lysosome Mitochondria Nucleus Peroxisome Plasma membrane
1.5 1 0.5 0
B
-log10(p-value) acetylation changes 24 hpi
Cytosol 0.21 0.03 0.00 0.50 0.00 0.00 1.07 0.00 0.22
Dense cytosol 1.39 0.09 0.21 0.75 0.00 0.00 0.02 0.00 0.35
ER 0.00 0.33 0.37 0.24 0.66 0.00 0.00 0.00 0.00
Golgi 0.00 0.00 0.97 0.28 0.00 0.00 0.00 0.00 0.00
Lysosome 0.00 0.00 0.38 0.00 0.38 0.00 0.00 0.00 0.00
Mitochondria 0.00 0.00 0.00 0.00 0.00 0.31 0.00 0.00 0.00
Nucleus 0.00 0.01 0.00 0.00 0.00 0.33 0.00 0.28 0.00
Peroxisome 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
Plasma membrane 0.24 0.25 0.00 0.00 0.00 0.00 0.00 0.00 0.49
Cytosol Dense cytosol ER Golgi Lysosome Mitochondria Nucleus Peroxisome Plasma membrane
2 1.5 1 0.5 0
C
Cytosol 167 21 0 7 0 0 6 0 1
Dense cytosol 34 79 79 47 0 1 0 1 26
ER 0 14 87 56 5 0 0 0 0
Golgi 0 0 7 68 4 0 0 0 0
Lysosome 0 0 42 4 33 0 0 0 0
Mitochondria 0 0 0 0 0 309 0 5 0
Nucleus 0 13 0 14 0 31 7 15 0
Peroxisome 0 0 0 0 0 0 0 29 0
Plasma membrane 2 8 0 4 0 0 0 0 206
Cytosol Dense cytosol ER Golgi Lysosome Mitochondria Nucleus Peroxisome Plasma membrane
300 250 200 150 100 50 0
D
UL148A interactome
Cytosol 3 0 0 0 0 0
Dense cytosol 10 7 3 0 0 1
ER 0 0 1 0 0 0
Golgi 0 3 2 1 0 0
Lysosome 0 2 1 2 0 0
Mitochondria 0 0 0 0 1 0
Plasma membrane 0 0 1 0 0 8
Cytosol ER Golgi Lysosome Mitochondria Plasma membrane
10 8 6 4 2 0
E
UL70 interactome
Cytosol 7 1 0 0 0
Dense cytosol 0 1 0 0 0
ER 0 1 0 0 0
Golgi 0 0 1 0 0
Mitochondria 0 0 0 1 0
Plasma membrane 0 0 0 0 1
Cytosol Dense cytosol ER Mitochondria Plasma membrane
7 6 5 4 3 2 1 0
F
UL8 interactome
Dense cytosol 4 1 0 0
Golgi 0 0 1 0
Lysosome 0 1 0 0
Plasma membrane 0 0 0 7
Cytosol ER Lysosome Plasma membrane
7 6 5 4 3 2 1 0

Figure 7:

Figure 7: (A) A heatmap representation of the mean log2 fold changes in acetylation overlaid on spatial pattern of HCMV infection 24 hpi. (B) p-values for the changes shown in figure A. (C) The spatial allocation derived from BANDLE where each entry of the heatmap is the number of proteins. The y-axis represents localisation in the mock dataset and the x-axis localisation in the HCMV infected cells 24 hpi. (D) UL148A interactome mapped onto the BANDLE determined spatial patterns. (E) UL70 interactome mapped onto the BANDLE determined spatial patterns (F) UL8 interactome mapped onto the BANDLE determine spatial patterns.

The recent publication of the HCMV interactome has provided a wealth of data that gives insights into the function of the 170 canonical and 2 non-canonical viral protein-coding genes (Nobre *et al.*, 2019). However, a common difficulty with analysing large interactome projects is the ability to reduce the number of false-positive interactions, leading to poor agreement between experimental and computational datasets. This can be controlled through replicates, supervised machine learning and increased statistical stringency; however, background contamination can never be eliminated. If a protein is located in a single location, you would expect true positive interactors to be located in the same subcellular compartment. Therefore, to narrow the list of viral-protein interactors, we overlapped spatial information from Beltran *et al.* (2016) with the viral interactors from Nobre *et al.* (2019) (figure 7 D,E,F).

We plot heatmaps to indicate the spatial distribution of the host proteins (figure 7). The overall distribution is plotted in the heatmap of figure 7 C. Firstly, we are interested in scenarios where the interacting host proteins were more likely to retain their localisation upon HCMV infection (than the computed posterior distribution would have predicted). Thus, for each viral bait, we simulated from a binomial $A \sim Bin(n, p)$ where p is the posterior probability that a random protein was assigned to the same localisation and n is the number of interactors of that viral bait. We then simulated from this distribution $5,000$ times to obtain a histogram (see supplement). Viral baits of interest are those were the observed statistic in the tails of these histograms.

Examples of such cases are shown for viral proteins UL8 and UL70 (see figure 7 E and F). The majority of UL8 interactors were located in the plasma membrane and cytosol. UL8 is a transmembrane protein that is transiently localised at the cell surface, with a small cytoplasmic pool (Pérez-Carmona *et al.*, 2018), perfectly mimicking the location of the majority of UL8 interactors. Practically all UL70 interactors were located in the cytosol. Viral UL70 is a primase known to locate to both the nucleus and cytoplasmic compartments during HCMV infection (Shen *et al.*, 2011). As the nucleus was removed prior to fractionation then one expects only to be able to interrogate cytosolic interactors. An example were the host proteins were spatially diffuse was UL148A an elusive viral protein of unknown function, believed to be involved with modulating the innate immune response (Dassa *et al.*, 2018). UL148A appears to interact with host proteins distributed throughout the cell suggesting it is highly promiscuous (figure 7 D). Perhaps UL148A is a moonlighting protein (Jeffery, 2009) making its function hard to pinpoint and such an observation would not be uncommon for viral proteins because of limited genomic size (Cook and Lee, 2013; Copley, 2014). These results illustrate the strength in overlapping spatial proteomics with interactome studies to decrease the number of false positives and focus research on higher confidence protein-protein interactions. The entire list of spatially resolved viral protein interactions is shown in the supplementary material.

3 Discussion

We have presented a Bayesian model for comparative and dynamic spatial proteomic experiments. Unlike current approaches, our flexible integrative mixture model allows any number of replicate experiments to be included. Furthermore, subcellular profiles are modelled separately for each condition and each replicate, allowing cases where the correlation profiles differ between experiments. Crucially, our model facilitates the computation of differential localisation probability, which cannot be performed by other methods in the literature. Furthermore, BANDLE probabilistically assigns proteins to organelles and can model outliers meaning that further supervised machine learning after application of BANDLE is not required. The probabilistic ranking obtained from BANDLE can be used for downstream pathway or GO enrichment analysis, likewise it can be mapped onto other orthogonal high-throughput datasets.

We compared BANDLE to the MR approach of (Itzhak *et al.*, 2016, 2017). The MR method is not as broadly applicable as BANDLE, and BANDLE does not require additional experiments to interpret the thresholds. In our careful simulation study, we demonstrate reduced Type 1 error and increased power when using our approach. In a further simulation, we demonstrated that BANDLE has more desirable statistical properties than the MR approach, the results are easier to interpret and more information is available. Since we are in a Bayesian framework, our approach also quantifies uncertainty.

Application of our approach to 3 dynamic and comparative mass-spectrometry based spatial proteomic experiments demonstrates the broad applicability of our approach. We validate many previously known finding in the literature, placing confidence in these results. When BANDLE was applied to EGF stimulation dataset, we saw increase correlation between our differential localisation results and a phosphoproteomic timecourse than when compared to the results of the MR approach.

We applied BANDLE to an AP-4 knockout dataset to investigate AP-4 dependant localisation and, as with other studies, we observe SERINC 1 and SERINC 3 are AP-4 Cargo. Furthermore, we implicate TMEM199 as potentially overlooked AP-4 cargo. We apply BANDLE to a datasets where the MR approach is not applicable - an HCMV infection spatial proteomic dataset. Pathway and GO enrichment results implicate differentially localised protein in well-studied processes of early viral infection; such as, membrane trafficking and immune response.

We then carefully integrated several HCMV proteomic datasets and place a spatial perspective on these data, including proteins targeted for degradation, as well as abundance and acetylation dataset. In addition, we augment a recent HCMV interactome by placing it in its spatial context and note that most host protein interactomes are in the same localisation as their viral bait. This provides an excellent resource for the community and highlights the benefit of integrating spatial proteomics and interactomics datasets. This analysis also reveals potential moonlighting proteins.

Our analysis here highlights the potential role for post-translational modifications (PTMs) and their influence on localisation. The current datasets are limited because the spatial information is averaged over different PTMs. Thus, it is vital to develop methods to obtain spatial PTM information and develop corresponding computational tools to analyse these data. Furthermore, our approach here can only look at a single condition at a time. In the future, more complex spatial proteomics designs will be available that will study multiple perturbations simultaneously. This is the subject of future work.

Overall, differential localisation experiments seek to add an orthogonal perspective to other

assays, such as classical high-throughput differential abundance testing. Currently, differential localisation has not been extensively explored in high-throughput. We hope rigorous statistical methods will spur extensive and illuminating applications. An R-package is provided for analysis at `https://github.com/ococrook/bandle`, building on a suite of packages for analysing spatial proteomics data Gatto and Lilley (2012); Gatto *et al.* (2014b); Crook *et al.* (2019a).

4 Methods

4.1 The Movement-Reproducibility method

The movement-reproducibility (MR) method was proposed by Itzhak *et al.* (2016, 2017) and this is our interpretation of their method. We suppose that we are given two spatial proteomics experiments under a single contrast/perturbation/treatment, and denote unperturbed by $(d = 1)$ and $(d = 2)$ for the perturbed condition. Furthermore, assume we measure each condition with $r = 1, ..., R$ biological replicates. Let $X_1 = [X_1^{(1)}, ..., X_1^{(R)}]$ denote the concatenation of replicates for condition 1 and likewise for condition 2 denotes $X_2 = [X_2^{(1)}, ..., X_2^{(R)}]$. We first compute delta matrices as follows

$$\Delta = X_1 - X_2, \tag{1}$$

where $\Delta = [\Delta^{(1)}, ..., \Delta^{(R)}]$. This assumes that both features and replicates are comparable in some way; that is, a feature in the rth replicate is directly comparable to the same feature in another replicate. Then for each Δ_r, $r = 1, .., R$, the squared Mahalanobis distance D_M from each protein to the empirical mean is computed using a robust estimate of the covariance matrix - the minimum covariance determination method (Hubert and Debruyne, 2010). Under a Gaussian assumption on Δ_r, $D_M(p_i)$ follows a chi-squared distribution with degrees of freedom equal to the dimension of the data G. Then, for each protein and each replicate a p-value is computed, such that there are R such p-values for each protein. These p-values are combined into a score by taking the cube of the largest p-value for each protein, correcting for multiple hypothesis testing using the Benjamini-Höchberg procedure and computing the $-\log_{10}$ of the resultant value. The final score is called the M score.

This process means that the computed value can no longer be interpreted as truly derived from a p-value. To maintain this interpretation one could instead combine p-values using Fisher's method (Mosteller and Fisher, 1948). Furthermore, the authors are, implicitly, concerned with finding *any* false positives and as such control over the FWER is desired rather than the FDR. Since FWER $\geq$ FDR, control of the FDR does not lead to control over the FWER.

A so-called reproducibility (R) score is obtained by first computing the pearson correlation pairwise between matrices $\Delta_i, \Delta_j, i \neq j$ for each protein. A final R score, for each protein, is obtain by taking the minimum value for each protein. Again this score could have be interpret in a formal testing procedure using a permutation test (Efron, 2012) and furthermore includes an assumption of bivariate normality. Moreover, Pearson's correlation is unresponsive to many non-linear relationships which might be present.

Finally, each protein has an associated pair of scores, referred to as the MR-score. To determine thresholds for these scores the authors take a desired FDR $= 0.01$. Thus they repeat a control experiment 6 times to determine thresholds $M = 2, R = 0.9$ a region with no false discoveries.

Repeating the control experiment 6 times is a costly process and likely to be prohibitive for most

experiments, particularly for cells that are expensive to culture. Furthermore, since the thresholds are empirically derived, this process needs to be repeated for every new experiment to determine optimal thresholds.

4.2 BANDLE

4.2.1 A model for differential localisation

In the following, we layout our model for BANDLE, along with methods for inference, and approaches for summarising and visualising the output. Firstly, suppose we have two spatial proteomics experiments with unperturbed ($d = 1$) and perturbed conditions ($d = 2$). Furthermore, assume we measure each condition with $r = 1, ..., R$ biological replicates. Let $X_1 = [X_1^{(1)}, ..., X_1^{(R)}]$ denote the concatenation of replicates for condition 1 and likewise for condition 2 denotes $X_2 = [X_2^{(1)}, ..., X_2^{(R)}]$. We introduce the following latent allocation variable $z_{i,d}$, denoting the localisation of protein i in condition d. Thus, if $z_{i,d} = k$ this means that protein i localises to organelle k in dataset d. Given this latent allocation variable, we assume that the data from replicate $r = 1, ..., R$ arises from some component density $F(\cdot|\theta_k^{(r)})$. Hence, denoting by θ the set of all component parameters, we can write

$$x_{i,d}^{(r)}|z_{i,d}, \theta \sim F(x_{i,d}^{(r)}|\theta_{z_{i,d}}^{(r)}). \tag{2}$$

We assume that biological replicates are independent and so we factorise as follows

$$p(x_{i,d}|z_{i,d}, \theta) = \prod_{r=1}^{R} p(x_{i,d}^{(r)}|z_{i,d}, \theta_{z_{i,d}}^{(r)}). \tag{3}$$

To couple the two conditions together we assume a joint prior structure for the latent allocation variable in each dataset. To be more precise, we construct a prior for the pair $(z_{i,1}, z_{i,2})$. We fix the possible number of subcellular niches to which a protein may localise to be K. Now, we introduce the matrix Dirichlet distribution, which we denote as $\mathcal{M}\text{Dir}(\alpha, K)$. The concentration parameter α is a $K \times K$ matrix, such that for a matrix $\boldsymbol{\pi}$, the pdf of the matrix Dirichlet distribution is

$$f(\boldsymbol{\pi}|\alpha) = \prod_{k=1}^{K} \frac{1}{\mathcal{B}(\alpha_k)} \prod_{j=1}^{K} \pi_{jk}^{\alpha_{jk}-1}, \tag{4}$$

where $\mathcal{B}$ denotes the beta function, α_k denotes the kth row of α, and $\sum_{j,k} \pi_{jk} = 1$. Thus, we propose the following hierarchical structure

$$\boldsymbol{\pi}|\alpha \sim \mathcal{M}\text{Dir}(\alpha, K) \tag{5}$$

$$(z_{i,1}, z_{i,2}) \sim cat(\boldsymbol{\pi}), \tag{6}$$

where $(z_{i,1}, z_{i,2}) \sim cat(\boldsymbol{\pi})$ means that the prior allocation probabilities are given by

$$p(z_{i,1} = k, z_{i,2} = k'|\boldsymbol{\pi}) = \pi_{kk'}. \tag{7}$$

The above model is conjugate, and so if $n_{j,k} = |\{(z_{i,1}, z_{i,2}) = (j, k)\}|$, it follows that the conditional posterior of $\boldsymbol{\pi}$ is

$$\boldsymbol{\pi}|(Z_1, Z_2), \alpha \sim \mathcal{M}\text{Dir}(\gamma, K) \tag{8}$$

where $\gamma_{j,k} = \alpha_{jk} + n_{j,k}$. The likelihood models for the data are Gaussian Random Fields, which we elaborate on in the following section. Hence, the conditional posterior of the allocation probabilities are

$$p(z_{i,1} = j, z_{i,2} = k|\boldsymbol{\pi}) \propto \pi_{jk} \prod_{r=1}^{R} p(x_{i,1}^{(r)}|z_{i,1} = j)p(x_{i,2}^{(r)}|z_{i,2} = k). \tag{9}$$

4.2.2 Likelihood Model

The model described in the previous section is presented in a general form, so it could be applied to many different modes of data. We describe the model for a single spatial proteomics experiment, since the same model is assumed across all spatial proteomics experiments, that are then subsequently joined together using the approach in the previous section. Though the model is the same across experiments, the parameters are experiment-specific.

We assume that the protein intensity x_i at each fraction s_j can be described by some regression model with unknown regression function:

$$x_i(s_j) = \mu_i(s_j) + \varepsilon_{ij}, \tag{10}$$

where μ_i is some unknown deterministic function of space and ε_{ij} is a noise variable, which we assume is $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_i^2)$. Proteins are grouped together according to their subcellular localisation; such that, all proteins associated to subcellular niche $k = 1, ..., K$ share the same regression model. Hence, we write $\mu_i = \mu_k$ and $\sigma_i = \sigma_k$. Throughout, for clarity, we refer to sub-cellular structures, whether they are organelles, vesicles or large protein complexes, as *components*. The regression functions μ_k are unknown and thus we place priors over these functions to represent our prior uncertainty. Protein intensities are spatially correlated and thus we place *Gaussian Random Field* (GRF) priors over these regression functions. We pedantically refer to these as GRF priors rather than *Gaussian Process* (GP) priors to make the distinction between the 1D spatial process that separates sub-cellular niches and the experimental cellular perturbations, which are potentially temporal in nature. Hence, we write the following

$$\mu_k \sim GRF(m_k(\boldsymbol{s}), C_k(\boldsymbol{s}, \boldsymbol{s'})), \tag{11}$$

which is defined as:

Definition 1. *Gaussian Random Field*
If $\mu(\boldsymbol{s}) \sim GRF(m_k(\boldsymbol{s}), C_k(\boldsymbol{s}, \boldsymbol{s'}))$ then for any finite dimensional collection of indices $s_1, ..., s_n$, $[\mu(s_1), ..., \mu(s_n)]$ is multivariate Gaussian with mean $[m(s_1), ..., m(s_n)]$ and covariance matrix such that $C_{ij} = C(s_i, s_j)$.

Each component is thus captured by a Gaussian Random Field model and the full complement of proteins as a finite mixture of GRF models. The protein intensity for each experiment maybe measured in replicate. For a sufficiently flexible model, we allow different regression models across different replicates. To be more precise, consider the protein intensity $x_i^{(r)}$ for the ith protein measured in replicate r at fraction $s_j^{(r)}$, then we can write the following

$$x_i^{(r)}\left(s_j^{(r)}\right) = \mu_k^{(r)}\left(s_j^{(r)}\right) + \varepsilon_{ij}^{(r)}, \tag{12}$$

having assumed that the ith protein is associated to the kth component. The (hyper)parameters for the Gaussian Random Field priors for the rth replicate in experiment d are denoted by $\theta_{k,d}^{(r)}$. We denote by $\boldsymbol{\theta}$ the collection of all hyperparameters and the collection of priors for these by $G_0(\boldsymbol{\theta})$. The loss of conjugacy between the prior on the hyperparameters and likelihood is unavoidable.

The GRF is used to model the uncertainty in the underlying regression functions; however, we have yet to consider the uncertainty that a protein belongs to each of these components. To capture these uncertainties, we can use the model in the previous section, allowing information to be shared across each condition. Following from the previous section, the conditional posterior of the allocation probabilities is

$$p(z_{i,1} = j, z_{i,2} = k|\boldsymbol{\pi}) \propto \pi_{jk} \prod_{r=1}^{R} p(x_{i,1}^{(r)}|z_{i,1} = j)p(x_{i,2}^{(r)}|z_{i,2} = k), \tag{13}$$

where, in the specific case of our likelihood model the probabilities in the terms of the product can be computed using the appropriate GRF.

We assume that our Gaussian processes are centred and that the covariance is from the Matern class (Stein, 1999). The Matern covariance is specified as follows

$$C_v(d) = a^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{8\nu}\frac{d}{\rho}\right)^{\nu} \mathcal{K}_v \left(\sqrt{8\nu}\frac{d}{\rho}\right), \tag{14}$$

where Γ is the gamma function and $\mathcal{K}_v$ denotes the modified Bessel function of the second kind of order $\nu > 0$. Furthermore, a and ρ are positive parameters of the covariance. a^2 is interpreted as a marginal variance, whilst the non-standard choice of $\sqrt{8\nu}$ in the definition of the Matern covariance, allows us to interpret ρ as a range parameter and thus ρ is the distance at which the correlation is 0.1 for any ν (Lindgren *et al.*, 2011). The Matern covariance arises from solutions of the following linear fractional stochastic partial differential equation (SPDE):

$$(\kappa^2 - \Delta)^{\alpha/2} x(u) = \mathcal{W}(u), u \in \mathbb{R}^d \, \alpha = \nu + d/2, \;\; \kappa > 0, \nu > 0. \tag{15}$$

where $\mathcal{W}(u)$ is spatial Gaussian white noise with unit variance and Δ is the Laplacian. The parameter ν controls the differentiability of the resulting sample paths; such that, $\lceil v \rceil$ is the number of mean-square derivatives. For typical applications, ν is poorly identifiable and fixed. $\nu = 1/2$ recovers the exponential covariance, whereas taking the limit $\nu \to \infty$ one obtains the squared exponential (Gaussian) covariance. We fix $\nu = 2$.

A ridge in the marginal likelihood for the marginal variance and range parameters of the Matern covariance makes inference challenging. Indeed, different hyperparameters lead to unconditional prior simulations with the same spatial pattern but different scales (Rasmussen and Williams, 2006; Fuglstad *et al.*, 2019). Furthermore, when the intrinsic dimension of the Gaussian random field is less than four, there is no consistent estimator under in-fill asymptotics for ρ and a. A principled prior, which allows domain expertise to be expressed, is thus desired to enable stable inferences. A number of works considered reference priors for GRFs (Berger *et al.*, 2001; Paulo *et al.*, 2005; De Oliveira, 2007; van der Vaart *et al.*, 2009). Here, we employ a recently introduced collection of weakly-informative priors, which we introduce in the supplementary methods.

4.2.3 Calibration of Dirichlet prior

The following section describes how to calibrate the prior based on expert information and prior predictive checks. Recall the prior on the allocation probabilities is the following

$$p(z_{i,1} = k, z_{i,2} = k' | \boldsymbol{\pi}) = \pi_{kk'}. \tag{16}$$

The matrix $\boldsymbol{\pi}$ has π_{jk} has its $(j,k)^{th}$ entry and π_{jk} is the prior probability that a protein belongs to organelle j in dataset 1 (control) and k in dataset 2 (contrast). The diagonal terms represent the probability that the protein was allocated to the same organelle in each dataset. The non-diagonal terms are the prior probability that the protein was not allocated to the same organelle. Since the number of non-diagonal terms greatly exceeds to the number of diagonal entries it is important to specify this prior carefully. Recall that the prior is given a matrix Dirichlet distribution with concentration parameter α.

Firstly, we are interested in the prior expectation of the number of protein that are *differential localised*; that is, proteins not allocated to the same organelle in both conditions. Let ρ be the prior probability that a protein is not allocated to the same organelle. Then it follows that

$$p(z_{i,1} \neq z_{i,2} | \boldsymbol{\pi}) =: \rho = \sum_{j,k;j\neq k} \pi_{jk} \tag{17}$$

By properties of the Dirichlet distribution we have that

$$\pi_{jk} \sim \mathcal{B}(\alpha_{jk}, \alpha_0 - \alpha_{jk}). \tag{18}$$

Thus, the expected value of ρ is computed as follows

$$\begin{aligned} \mathbb{E}[\rho] &= \sum_{j,k;j\neq k} \mathbb{E}[\pi_{jk}] \\ &= \sum_{j,k;j\neq k} \frac{\alpha_{jk}}{\alpha_0}. \end{aligned} \tag{19}$$

We are further interested in the probability that a certain number of proteins, say q, are differential localised. Letting N_U be the number of unlabelled proteins in the experiment, then the distribution of the prior number of differential localised proteins is

$$p(N_U \rho > q) = p\left(N_U \sum_{j,k;j\neq k} \pi_{jk} > q \right) = \delta \tag{20}$$

Computing δ is not simple; however, it is straightforward to estimate δ using Monte-Carlo by simply sampling from Beta distributions:

$$p\left(N_U \sum_{j,k;j\neq k} \pi_{jk} > q \right) \approx \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\left(N_U \sum_{j,k;j\neq k} \pi_{jk}^{(t)} > q \right). \tag{21}$$

Thus, we recommend calibrating the Dirichlet prior using the above expectation and quantile. It many be important to calibrate several quantiles to ensure sufficient mass is placed on desired regions of the probability space. For example, let $q_1 < q_2$, then we may desire that δ_1, below, is

not so small to rule out reasonable inferences and that $\delta_2 < \delta_1$ is sufficiently large. These can be computed from the equations below.

$$p\left(N_U \sum_{j,k;j\neq k} \pi_{jk} > q\right) \approx \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\left(N_U \sum_{j,k;j\neq k} \pi_{jk}^{(t)} > q_1\right) = \delta_1, \tag{22}$$

$$p\left(N_U \sum_{j,k;j\neq k} \pi_{jk} > q\right) \approx \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\left(N_U \sum_{j,k;j\neq k} \pi_{jk}^{(t)} > q_2\right) = \delta_2. \tag{23}$$

More precise and informative prior biological knowledge can be specified; for example, should we suspect that some relocalisation events between particular organelles are more likely than others due to the stimuli, these can be encode into the prior. If we expect more relocalisation events between organelle j and k_1 than organelle j and k_2, this can be encoded by ensuring

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\left(\pi_{jk_1}^{(t)} > \pi_{jk_2}^{(t)}\right) > \delta_3 > 0. \tag{24}$$

Alternatively, if an objective Bayesian analysis is preferred, the Jeffery's prior sets $\alpha_{jk} = 0.5$ for every $j, k = 1, .., K$. This approach is not generally recommended by the authors, because the diagonal terms of $\boldsymbol{\pi}$ have a different interpretation to the off-diagonal terms.

4.2.4 Differential localisation probability

The main posterior quantity of interest is the probability that a protein is differentially localised. This can be approximated from the T Monte-Carlo samples as follows, suppressing notational dependence on all data and parameters for clarity

$$\chi_i = p(z_{i,1} \neq z_{i,2}) \approx \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}(z_{i,1}^{(t)} \neq z_{i,2}^{(t)}), \tag{25}$$

where t denotes the tth sample of the MCMC algorithm. It is important to note that this quantity is agnostic to the assigned subcellular niche.

To perform uncertainty quantification on the the differential localisation probability, we use the non-parametric bootstrap on the Monte-Carlo samples. More precisely, first sample *uniformly with replacement* from $\{z_{i,1}^{(t)}\}_{t=1}^{T}$ and $\{z_{i,2}^{(t)}\}_{t=1}^{T}$ to total of T samples. This produces a bootstrap sample indexed by B_1. Then compute our statistic of interest

$$\chi_{i,B_1}^{*} \approx \frac{1}{|B_1|}\sum_{t\in B_1} \mathbb{1}(z_{i,1}^{(t)} \neq z_{i,2}^{(t)}). \tag{26}$$

This process is then repeated to obtain a set of bootstrap samples $\mathbb{B} = \{B_1, ..., B_b\}$, for some large b, say 1000. For each $B_r \in \mathbb{B}$, we compute χ_{i,B_r}^{*} for $r = 1, ..., b$, obtaining a sampling distribution for χ_r from which we can compute functionals of interest.

4.2.5 Posterior localisation probabilities

A further quantity of interest is the posterior probability that a protein belongs to each of the K sub-cellular niches present in the data. For the control this is given by the following Monte-Carlo average

$$p(z_{i,1} = k|\Theta) \approx \frac{1}{T}\sum_{t=1}^{T} p(z_{i,1}^{(t)} = k|\Theta), \tag{27}$$

where Θ denotes all other quantities in the model. A corresponding formula also holds for the second dataset

$$p(z_{i,2} = k|\Theta) \approx \frac{1}{T}\sum_{t=1}^{T} p(z_{i,2}^{(t)} = k|\Theta), \tag{28}$$

The posterior distribution of these quantities and uncertainty estimates can be computed and visualised in standard ways.

5 Funding

OMC is a Wellcome Trust Mathematical Genomics and Medicine student funded by the Cambridge School of Clinical Medicine. As well as MRC awards to PDWK MC_UU_00002/13 and MC_UU_00002/10. KSL is supported by the Wellcome Trust 110071/Z/15/Z. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. This work was supported by the National Institute for Health Research [Cambridge Biomedical Research Centre at the Cambridge University Hospitals NHS Foundation Trust] [*]. *The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

6 Competing Interest

Colin T.R. Davies is an employee of AstraZenca (AZ). AZ had no role in the study design, data collection and analysis, decision to publish, or preperation of the manuscript. All other author declare no competing interests.

7 Data availability

All spatial proteomics data is available in the Bioconductor package pRolocdata. All additional data are given referenced manuscripts and copies are provided as part of the supplementary material. MCMC files are provided at Zenodo: https://doi.org/10.5281/zenodo.4415369

8 Code availability

An R-package is provided at https://github.com/ococrook/bandle.

References

Alwine, J. C. (2012). The human cytomegalovirus assembly compartment: a masterpiece of viral manipulation of cellular processes that facilitates assembly and egress. *PLoS pathogens*, **8**(9).

Baers, L. L. et al. (2019). Proteome mapping of a cyanobacterium reveals distinct compartment organisation and cell-dispersed metabolism. *Plant Physiology*.

Balta, E.-A. et al. (2018). Phosphorylation modulates the subcellular localization of sox11. *Frontiers in molecular neuroscience*, **11**, 211.

Banerjee, A. et al. (2013). Bayesian learning of joint distributions of objects. In *Artificial Intelligence and Statistics*, pages 1–9.

Barylyuk, K. et al. (2020). A subcellular atlas of toxoplasma reveals the functional context of the proteome. *bioRxiv*.

Behne, R. et al. (2020). Adaptor protein complex 4 deficiency: a paradigm of childhood-onset hereditary spastic paraplegia caused by defective protein trafficking. *Human Molecular Genetics*, **29**(2), 320–334.

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.

Berger, J. O. et al. (2001). Objective bayesian analysis of spatially correlated data. *Journal of the American Statistical Association*, **96**(456), 1361–1374.

Bevilacqua, E. et al. (2005). Snat2 silencing prevents the osmotic induction of transport system a and hinders cell recovery from hypertonic stress. *FEBS letters*, **579**(16), 3376–3380.

Boeckh, M. et al. (2004). The impact of cytomegalovirus serostatus of donor and recipient before hematopoietic stem cell transplantation in the era of antiviral prophylaxis and preemptive therapy. *Blood*, **103**(6), 2003–2008.

Bozidis, P. et al. (2010). Trafficking of ul37 proteins into mitochondrion-associated membranes during permissive human cytomegalovirus infection. *Journal of virology*, **84**(15), 7898–7903.

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

Breckels, L. M. et al. (2016a). A bioconductor workflow for processing and analysing spatial proteomics data. *F1000Research*, **5**.

Breckels, L. M. et al. (2016b). Learning from heterogeneous data sources: an application in spatial proteomics. *PLoS computational biology*, **12**(5), e1004920.

Brown, M. B. (1975). A method for combining non-independent, one-sided tests of significance. *Biometrics*, pages 987–992.

Buchkovich, N. J. et al. (2010). Role of the endoplasmic reticulum chaperone bip, sun domain proteins, and dynein in altering nuclear morphology during human cytomegalovirus infection. *Journal of virology*, **84**(14), 7005–7017.

Cannon, M. J. et al. (2010). Review of cytomegalovirus seroprevalence and demographic characteristics associated with infection. *Reviews in medical virology*, **20**(4), 202–213.

Choi, H. M. et al. (2013). The polya-gamma gibbs sampler for bayesian logistic regression is uniformly ergodic. *Electronic Journal of Statistics*, **7**, 2054–2064.

Christian, F. et al. (2016). The regulation of nf-κb subunits by phosphorylation. *Cells*, **5**(1), 12.

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

Cook, J. D. et al. (2013). The secret life of viral entry glycoproteins: moonlighting in immune evasion. *PLoS pathogens*, **9**(5).

Copley, S. D. (2014). An evolutionary perspective on protein moonlighting.

Cristea, I. M. et al. (2010). Human cytomegalovirus pul83 stimulates activity of the viral immediate-early promoter through its interaction with the cellular ifi16 protein. *Journal of virology*, **84**(15), 7803–7814.

Crook, O. et al. (2019a). A bioconductor workflow for the bayesian analysis of spatial proteomics [version 1; peer review: awaiting peer review]. *F1000Research*, **8**(446).

Crook, O. et al. (2020a). A semi-supervised bayesian approach for simultaneous protein sub-cellular localisation assignment and novelty detection. *bioRxiv*.

Crook, O. M. et al. (2018). A bayesian mixture modelling approach for spatial proteomics. *PLOS Computational Biology*, **14**(11), 1–29.

Crook, O. M. et al. (2019b). Semi-supervised non-parametric bayesian modelling of spatial proteomics. *arXiv preprint arXiv:1903.02909*.

Crook, O. M. et al. (2020b). Moving profiling spatial proteomics beyond discrete classification. *Proteomics*, page 1900392.

Crow, M. S. et al. (2016). Diverse mechanisms evolved by dna viruses to inhibit early host defenses. *Critical reviews in biochemistry and molecular biology*, **51**(6), 452–481.

Das, S. et al. (2011). Spatial relationships between markers for secretory and endosomal machinery in human cytomegalovirus-infected cells versus those in uninfected cells. *Journal of virology*, **85**(12), 5864–5879.

Das, S. et al. (2007). Three-dimensional structure of the human cytomegalovirus cytoplasmic virion assembly complex includes a reoriented secretory apparatus. *Journal of virology*, **81**(21), 11861–11869.

Dassa, L. et al. (2018). The human cytomegalovirus protein ul148a downregulates the nk cell-activating ligand mica to avoid nk cell attack. *Journal of virology*, **92**(17), e00162–18.

Davies, A. K. et al. (2018). Ap-4 vesicles contribute to spatial control of autophagy via rusc-dependent peripheral delivery of atg9a. *Nature communications*, **9**(1), 3958.

De Oliveira, V. (2007). Objective bayesian analysis of spatial data with measurement error. *Canadian Journal of Statistics*, **35**(2), 283–301.

De Pace, R. et al. (2018). Altered distribution of atg9a and accumulation of axonal aggregates in neurons from a mouse model of ap-4 deficiency syndrome. *PLoS genetics*, **14**(4), e1007363.

Dell'Angelica, E. C. et al. (1998). Association of the ap-3 adaptor complex with clathrin. *Science*, **280**(5362), 431–434.

Ebrahimi-Fakhari, D. et al. (2020). Defining the clinical, molecular and imaging spectrum of adaptor protein complex 4-associated hereditary spastic paraplegia. *Brain*, **143**(10), 2929–2944.

Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press.

Feng, X. et al. (2006). Human cytomegalovirus pus24 is a virion protein that functions very early in the replication cycle. *Journal of virology*, **80**(17), 8371–8378.

Ficarro, S. B. et al. (2002). Phosphoproteome analysis by mass spectrometry and its application to saccharomyces cerevisiae. *Nature biotechnology*, **20**(3), 301–305.

Fuglstad, G.-A. et al. (2019). Constructing priors that penalize the complexity of gaussian random fields. *Journal of the American Statistical Association*, **114**(525), 445–452.

Gabasova, E. et al. (2017). Clusternomics: Integrative context-dependent clustering for heterogeneous datasets. *PLoS computational biology*, **13**(10), e1005781.

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. (2019). Assessing sub-cellular resolution in spatial proteomics experiments. *Current opinion in chemical biology*, **48**, 123–149.

Geladaki, A. et al. (2019). Combining lopit with differential ultracentrifugation for high-resolution spatial proteomics. *Nature Communications*, **10**, 331.

Gelman, A. et al. (1995). *Bayesian Data Analysis*. Chapman & Hall, London.

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.

Gibson, W. (2008). Structure and formation of the cytomegalovirus virion. In *Human cytomegalovirus*, pages 187–204. Springer.

Gilks, W. R. et al. (1995). *Markov chain Monte Carlo in practice*. Chapman and Hall/CRC.

Gonzalez-Gonzalez, I. et al. (2005). Immunohistochemical localization of the amino acid transporter snat2 in the rat brain. *Neuroscience*, **130**(1), 61–73.

Good, I. (1958). Significance tests in parallel and in series. *Journal of the American Statistical Association*, **53**(284), 799–813.

Hatanaka, T. et al. (2000). Primary structure, functional characteristics and tissue expression pattern of human ata2, a subtype of amino acid transport system a. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, **1467**(1), 1–6.

Hess, J. et al. (2004). Ap-1 subunits: quarrel and harmony among siblings. *Journal of cell science*, **117**(25), 5965–5973.

Hirst, J. et al. (1999). Characterization of a fourth adaptor-related protein complex. *Molecular biology of the cell*, **10**(8), 2787–2802.

Hirst, J. et al. (2013). Adaptor protein complexes ap-4 and ap-5: new players in endosomal trafficking and progressive spastic paraplegia. *Traffic*, **14**(2), 153–164.

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

Holmes, S. et al. (2018). *Modern statistics for modern biology*. Cambridge University Press.

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

Hubert, M. et al. (2010). Minimum covariance determinant. *Wiley interdisciplinary reviews: Computational statistics*, **2**(1), 36–43.

Hwang, J. et al. (2007). Proteasome-dependent, ubiquitin-independent degradation of daxx by the viral pp71 protein in human cytomegalovirus-infected cells. *Virology*, **367**(2), 334–338.

Isaacson, M. et al. (2008). Virus entry and innate immune activation. In *Human Cytomegalovirus*, pages 85–100. Springer.

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.

Ivankovic, D. et al. (2020). Axonal autophagosome maturation defect through failure of atg9a sorting underpins pathology in ap-4 deficiency syndrome. *Autophagy*, **16**(3), 391–407.

Jansen, J. C. et al. (2016). Tmem199 deficiency is a disorder of golgi homeostasis characterized by elevated aminotransferases, alkaline phosphatase, and cholesterol and abnormal glycosylation. *The American Journal of Human Genetics*, **98**(2), 322–330.

Jean Beltran, P. M. et al. (2014). The life cycle and pathogenesis of human cytomegalovirus infection: lessons from proteomics. *Expert review of proteomics*, **11**(6), 697–711.

Jeffery, C. J. (2009). Moonlighting proteins - an update. *Molecular BioSystems*, **5**(4), 345–350.

Kalejta, R. (2008). Functions of human cytomegalovirus tegument proteins prior to immediate early gene expression. In *Human cytomegalovirus*, pages 101–115. Springer.

Karin, M. et al. (1997). Ap-1 function and regulation. *Current opinion in cell biology*, **9**(2), 240–246.

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

Kennedy, M. A. et al. (2020). Transpire: A computational pipeline to elucidate intracellular protein movements from spatial proteomics data sets. *Journal of the American Society for Mass Spectrometry*, **0**(0), null. PMID: 32401031.

Kirk, P. et al. (2012). Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, **28**(24), 3290–3297.

Köksal, A. S. et al. (2018). Synthesizing signaling pathways from temporal phosphoproteomic data. *Cell reports*, **24**(13), 3607–3618.

Kost, J. T. et al. (2002). Combining dependent p-values. *Statistics & Probability Letters*, **60**(2), 183–190.

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

Lee, A.-Y. et al. (2012). Protein kinase wnk3 regulates the neuronal splicing factor fox-1. *Proceedings of the National Academy of Sciences*, **109**(42), 16841–16846.

Li, T. et al. (2013). Human cytomegalovirus tegument protein pul83 inhibits ifi16-mediated dna sensing for immune evasion. *Cell host & microbe*, **14**(5), 591–599.

Linderman, S. et al. (2015). Dependent multinomial models made easy: Stick-breaking with the polya-gamma augmentation. In *Advances in Neural Information Processing Systems*, pages 3456–3464.

Lindgren, F. et al. (2011). An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**(4), 423–498.

Lock, E. F. et al. (2013). Bayesian consensus clustering. *Bioinformatics*, **29**(20), 2610–2616.

Lučin, P. et al. (2018). Cytomegaloviruses exploit recycling rab proteins in the sequential establishment of the assembly compartment. *Frontiers in Cell and Developmental Biology*, **6**.

Mahon, C. et al. (2014). Cullin e3 ligases and their rewiring by viral factors. *Biomolecules*, **4**(4), 897–930.

Mattera, R. et al. (2017). Ap-4 mediates export of atg9a from the trans-golgi network to promote autophagosome formation. *Proceedings of the National Academy of Sciences*, **114**(50), E10697–E10706.

Melone, M. et al. (2006). Localization of the na+-coupled neutral amino acid transporter 2 in the cerebral cortex. *Neuroscience*, **140**(1), 281–292.

Milbradt, J. et al. (2007). Cytomegaloviral proteins pul50 and pul53 are associated with the nuclear lamina and interact with cellular protein kinase c. *Journal of general virology*, **88**(10), 2642–2650.

Miles, A. L. et al. (2017). The vacuolar-atpase complex and assembly factors, tmem199 and ccdc115, control hif1α prolyl hydroxylation by regulating cellular iron levels. *Elife*, **6**, e22693.

Mitchell, D. P. et al. (2009). Human cytomegalovirus ul28 and ul29 open reading frames encode a spliced mrna and stimulate accumulation of immediate-early rnas. *Journal of virology*, **83**(19), 10187–10197.

Moorman, N. J. et al. (2010). A targeted spatial-temporal proteomics approach implicates multiple cellular trafficking pathways in human cytomegalovirus virion maturation. *Molecular & Cellular Proteomics*, **9**(5), 851–860.

Moreno-De-Luca, A. et al. (2011). Adaptor protein complex-4 (ap-4) deficiency causes a novel autosomal recessive cerebral palsy syndrome with microcephaly and intellectual disability. *Journal of medical genetics*, **48**(2), 141–144.

Moser, K. et al. (2006). Phosphoproteomic analysis of rat liver by high capacity imac and lc- ms/ms. *Journal of proteome research*, **5**(1), 98–104.

Mosteller, F. et al. (1948). Questions and answers. *The American Statistician*, **2**(5), 30–31.

Motley, A. et al. (2003). Clathrin-mediated endocytosis in ap-2–depleted cells. *The Journal of cell biology*, **162**(5), 909–918.

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

Murphy, E. et al. (2003). Reevaluation of human cytomegalovirus coding potential. *Proceedings of the National Academy of Sciences*, **100**(23), 13585–13590.

Murray, L. et al. (2018). Orchestration of protein acetylation as a toggle for cellular defense and virus replication. *Nature communications*, **9**(1), 1–17.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, **9**(1), 141–142.

Niemann, I. et al. (2014). Intracellular trafficking of the human cytomegalovirus-encoded 7-transmembrane protein homologs pus27 and pul78 during viral infection: a comparative analysis. *Viruses*, **6**(2), 661–682.

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

Nightingale, K. et al. (2018). High-definition analysis of host protein stability during human cytomegalovirus infection reveals antiviral factors and viral evasion mechanisms. *Cell host & microbe*, **24**(3), 447–460.

Nobre, L. V. et al. (2019). Human cytomegalovirus interactome analysis identifies degradation hubs, domain associations and viral protein functions. *eLife*, **8**, e49894.

Oda, K. et al. (2005). A comprehensive pathway map of epidermal growth factor receptor signaling. *Molecular systems biology*, **1**(1).

Ogawa-Goto, K. et al. (2003). Microtubule network facilitates nuclear targeting of human cytomegalovirus capsid. *Journal of virology*, **77**(15), 8541–8547.

Orre, L. M. et al. (2019). Subcellbarcode: Proteome-wide mapping of protein localization and relocalization. *Molecular Cell*, **73**(1), 166 – 182.

Paulo, R. et al. (2005). Default priors for gaussian processes. *The Annals of Statistics*, **33**(2), 556–582.

Pérez-Carmona, N. et al. (2018). A prominent role of the human cytomegalovirus ul8 glycoprotein in restraining proinflammatory cytokine production by myeloid cells at late times during infection. *Journal of virology*, **92**(9), e02229–17.

Plotnikov, A. et al. (2011). The mapk cascades: signaling components, nuclear roles and mechanisms of nuclear translocation. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, **1813**(9), 1619–1633.

Polson, N. G. et al. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, **108**(504), 1339–1349.

Possemato, A. P. et al. (2017). Multiplexed phosphoproteomic profiling using titanium dioxide and immunoaffinity enrichments reveals complementary phosphorylation events. *Journal of proteome research*, **16**(4), 1506–1514.

Puertollano, R. et al. (2018). The complex relationship between tfeb transcription factor phosphorylation and subcellular localization. *The EMBO journal*, **37**(11).

Rasmussen, C. E. et al. (2006). *Gaussian processes for machine learning*. MIT Press.

Reeves, M. et al. (2005). Latency, chromatin remodeling, and reactivation of human cytomegalovirus in the dendritic cells of healthy carriers. *Proceedings of the National Academy of Sciences*, **102**(11), 4140–4145.

Reinhardt, B. et al. (2006). Human cytomegalovirus-induced reduction of extracellular matrix proteins in vascular smooth muscle cell cultures: a pathomechanism in vasculopathies? *Journal of general virology*, **87**(10), 2849–2858.

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.

Robinson, M. S. (2015). Forty years of clathrin-coated vesicles. *Traffic*, **16**(12), 1210–1238.

Shen, A. et al. (2011). Human cytomegalovirus primase ul70 specifically interacts with cellular factor snapin. *Journal of virology*, **85**(22), 11732–11741.

Shin, J. J. et al. (2020). Spatial proteomics defines the content of trafficking vesicles captured by golgin tethers. *Nature communications*, **11**(1), 1–13.

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*.

Stamminger, T. et al. (2002). Open reading frame ul26 of human cytomegalovirus encodes a novel tegument protein that contains a strong transcriptional activation domain. *Journal of virology*, **76**(10), 4836–4847.

Stein, M. L. (1999). *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media.

Stern-Ginossar, N. et al. (2012). Decoding human cytomegalovirus. *Science*, **338**(6110), 1088–1093.

Sullivan, D. P. et al. (2018). Deep learning is combined with massive-scale citizen science to improve large-scale image classification. *Nature biotechnology*, **36**(9), 820.

Szklarczyk, D. et al. (2019). String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, **47**(D1), D607–D613.

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

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

van der Vaart, A. W. et al. (2009). Adaptive bayesian estimation using a gaussian random field with inverse gamma bandwidth. *The Annals of Statistics*, **37**(5B), 2655–2675.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372.

Weekes, M. P. et al. (2014). Quantitative temporal viromics: an approach to investigate host-pathogen interaction. *Cell*, **157**(6), 1460–1472.

Wilson, D. J. (2019). The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences*, **116**(4), 1195–1200.

Yu, Y. et al. (2011). Viral effects on metabolism: changes in glucose and glutamine utilization during human cytomegalovirus infection. *Trends in microbiology*, **19**(7), 360–367.

Yurochko, A. (2008). Human cytomegalovirus modulation of signal transduction. In *Human Cytomegalovirus*, pages 205–220. Springer.

Zeltzer, S. et al. (2018). Virus control of trafficking from sorting endosomes. *MBio*, **9**(4), e00683–18.