Probabilistic modelling of protein sub-cellular localisation

Oliver M Crook[1], Claire M Mulvey[1], Paul D. W. Kirk[2], Kathryn S Lilley[1], Laurent Gatto[3,*]

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

*laurent.gatto@uclouvain.be http://lgatto.github.io/about

Introduction

In biology, **localisation is function** - understanding the sub-cellular localisation of proteins is paramount to comprehend the context of their functions. The cellular sub-division allows cells to establish a range of distinct micro-environments (Figure 1), each favouring different biochemical reactions and interactions and, therefore, allowing each compartment to fulfil a particular functional role.

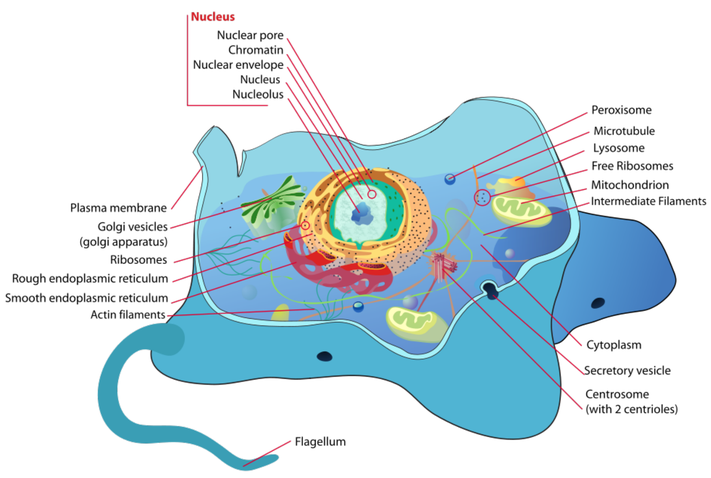

Figure 1 : Structure of an animal cell (credit: Wikipedia).

Mass spectrometry (MS)-based spatial proteomics (Figure 3) and contemporary machine learning enable to build **proteome-wide spatial maps** (Figure 2), informing us on the location of thousands of proteins.

Nevertheless, while some proteins can be found in a single location within a cell, up to half of proteins may reside in multiple locations, can dynamically re-localise, or reside within an unknown functional compartment, leading to considerable uncertainty in associating a protein to their sub-cellular location.

Recent advances enable to probabilistically model protein localisation as well as quantify the uncertainty in the location assignments, thus leading to better and more trustworthy biological interpretation of the data.

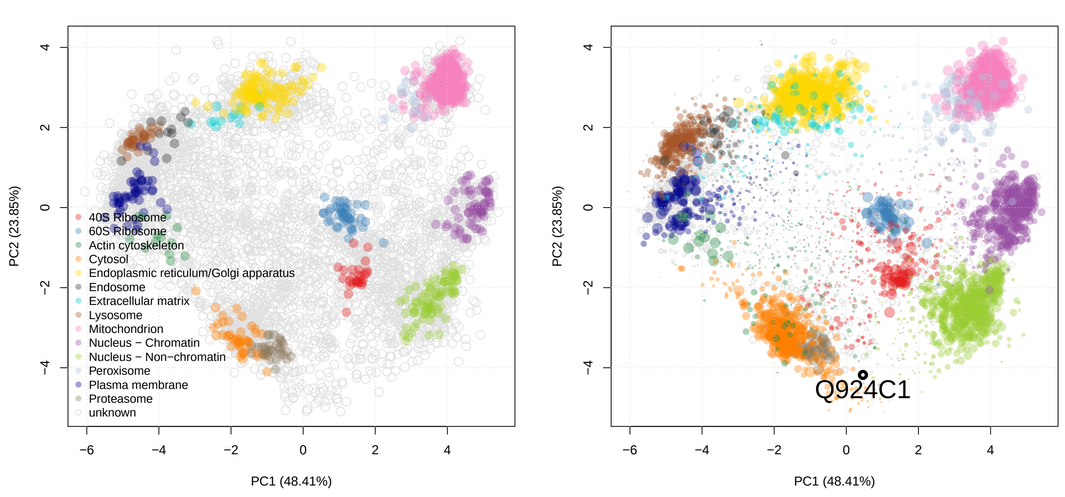

Figure 2 : Principle component analysis of the pluripotent mouse embryonic stem cell spatial map: each dot represents a single protein samples along 20 gradient fractions (Figure 3). Left: among the 5032 proteins, 926 marker proteins (well known and curated proteins that can be confidently assigned to a unique location) depicting 14 distinct and well resolved sub-cellular niches. Right: assignment of proteins of *unknown* location to one of the annotated classes. The dots are scaled according to the protein assignment probabilities. Note that while some sub-cellular clusters overlap along PC1 and PC2, they are separated along additional dimensions.

Mass spectrometry-based spatial proteomics

The *hyper*LOPIT protocol (Figure 3) uses density gradients to separate organelles and macro-molecular complexes. A set of discrete fractions are then collected and proteins are extracted, identified and quantified by mass spectrometry. The quantitaive proteins profiles display location specific patterns that are used for clustering and localisation analyses (classification) (Figure 2).

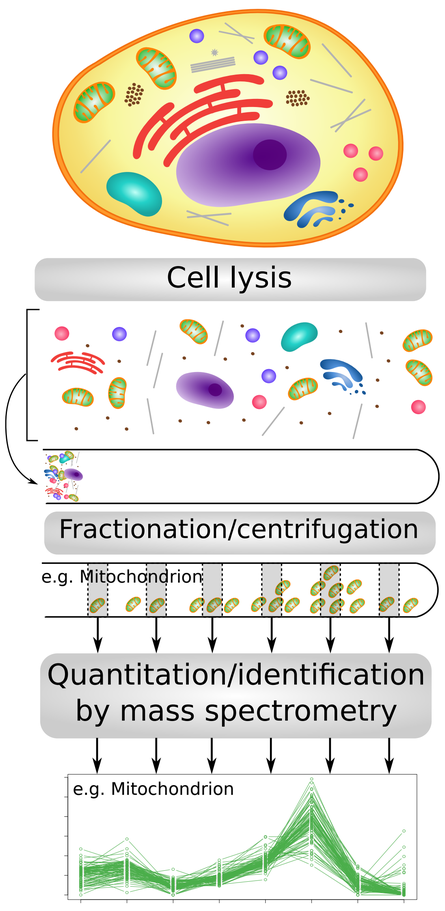

Figure 3 : LOPIT and *hyper*LOPIT experiments.

Funding This work was supported by the Wellcome Trust and the Biotechnology and Biological Sciences Research Council (BBSRC).

Results and conclusions

By implementing a probabilistic model for mass spectrometry-based spatial proteomics, we are in a position to deconvolute biologically important localisation patterns such as putative multi-localisation and confidently assign more proteins to their most likely sub-cellular localisations by quantifying uncertainty (Figure 4).

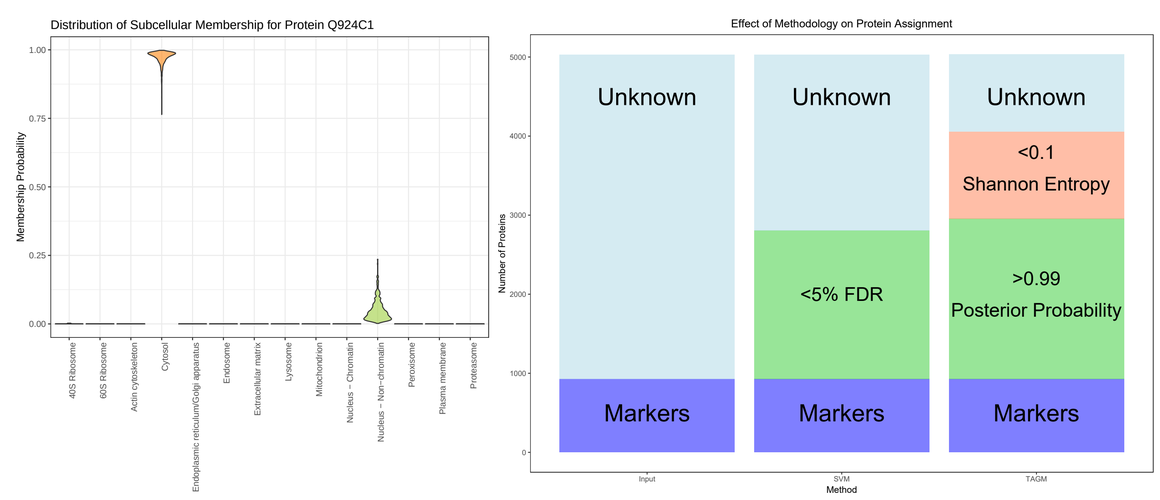

Figure 4 : Left: Exportin 5 (Q924C1, highlighted on Figure 2) forms part of the micro-RNA export machinery of the nucleus, transporting miRNA from the nucleus to the cytoplasm for further processing. It then translocates back through the nuclear pore complex to return to the nucleus. Exportin 5 can then continue to mediate further transport between nucleus and cytoplasm. Our model correclty infers that it most likely localises to the cytosol but there is some uncertainty with this assignment. This uncertainty is reflected in possible assignment of Exportin 5 to the nucleus non-chromatin and this uncertainty is a manifestation of the the fact that the function of this protein is to shuttle between the cytosol and nucleus.
Right: The barplot shows the number of marker proteins initially curated (left), those that are confidently assigned a unique localisation using a support vector machine classifier with a manually-assigned 5% false discovery threshold (center) and those that are assigned using our model with at least 99% posterior probability and low uncertainty (right). Roughly 2000 proteins are classified using either a support vector machine classifier and our TAGM model; however, TAGM can draw additional conclusions about an extra 1000 proteins by quantifying uncertainty using the Shannon entropy.

Probabilistic model

We present **T Augmented Gaussian Mixture model (TAGM)**, a multivariate Gaussian generative model for MS-based spatial proteomics data.

Our model posits that each annotated sub-cellular niche can be modelled by a multivariate Gaussian distribution. With the prior knowledge that many proteins are not captured by known sub-cellular niches, we augment our model with an outlier component. Outliers are often dispersed and thus this additional component is described by a heavy-tailed distribution: the multivariate Student's t-distribution, leading us to a T Augmented Gaussian Mixture model (TAGM) .

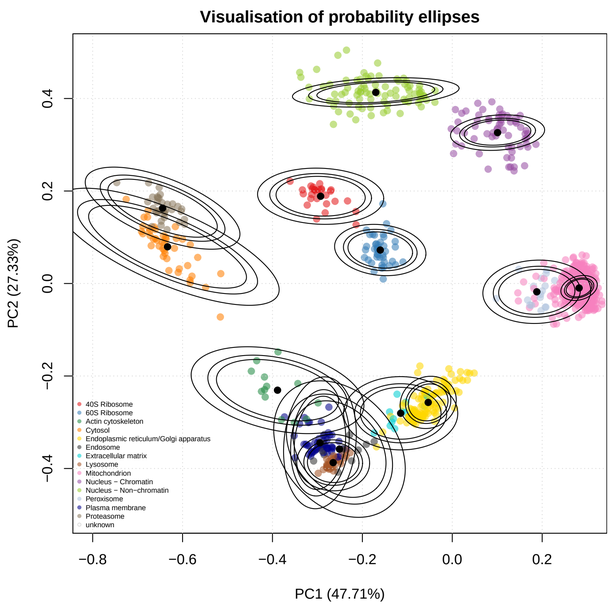

Figure 5 : Illustration of how the TAGM model describes the pluripotent mouse embryonic stem cell data. Each ellipse contains a proportion of total probability of a particular multivariate Gaussian density. The outer ellipse contains 99% of the total probability whilst the middle and inner ellipses contain 95% and 90% of the probability respectively.

References

A draft map of the mouse pluripotent stem cell spatial proteome Christoforou A *et al.* Nat Commun. 2016 Jan 12;7:8992. doi: 10.1038/ncomms9992.

A Bayesian Mixture Modelling Approach For Spatial Proteomics Crook OM *et al.* bioRxiv 282269; doi: 10.1101/282269

Software availablity

The TAGM algorithm is available as part of the `pRoloc` R/Bioconductor package. https://lgatto.github.io/pRoloc/.