Mapping the sub-cellular proteome
Probabilistic modelling of protein sub-cellular localisation

Laurent Gatto

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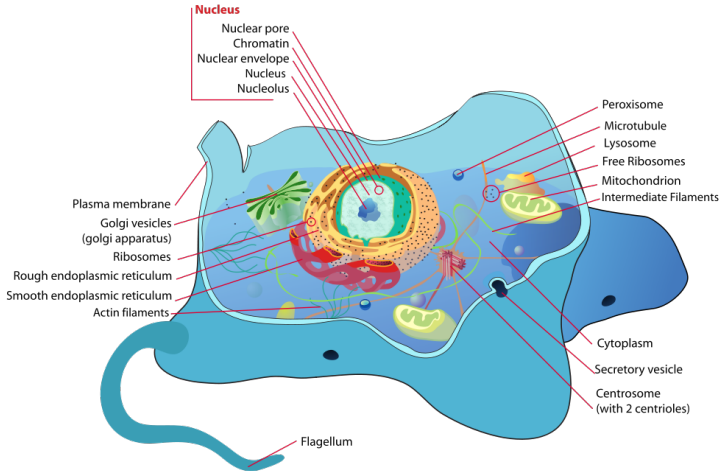
http://lgatto.github.io/about

de Duve Institute – UCLouvain

Protein Folding and Stability

30 August 2019 – Liège

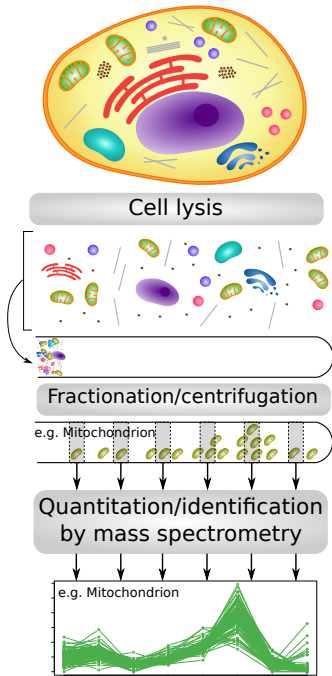# Cell organisation - **localisation is function**

Spatial proteomics is the systematic study of protein localisations.

Localisation – re-localisation – mis-localisation

Explorative/discovery approaches, steady-state **global localisation maps** (as opposed to microscopy-based targeted approaches).

Density gradient: PCP (Dunkley et al., 2006), LOPIT (Foster et al., 2006), hyperLOPIT (Christoforou et al., 2016; Mulvey et al., 2017) and **Differential centrifugation** Itzhak et al. (2016), LOPIT-DC (Geladaki et al., 2018).

Cell lysis

Fractionation/centrifugation

e.g. Mitochondrion

Quantitation/identification by mass spectrometry

e.g. Mitochondrion

Quantitation data

	Fraction$_1$	Fraction$_2$	. . .	Fraction$_L$
$\mathbf{x}_1$	$x_{1,1}$	$x_{1,2}$	. . .	$x_{1,L}$
$\mathbf{x}_2$	$x_{2,1}$	$x_{2,2}$	. . .	$x_{2,L}$
$\mathbf{x}_3$	$x_{3,1}$	$x_{3,2}$	. . .	$x_{3,L}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{x}_i$	$x_{i,1}$	$x_{i,2}$	. . .	$x_{i,L}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{x}_N$	$x_{N,1}$	$x_{N,2}$	. . .	$x_{N,L}$

Quantitation data and organelle markers

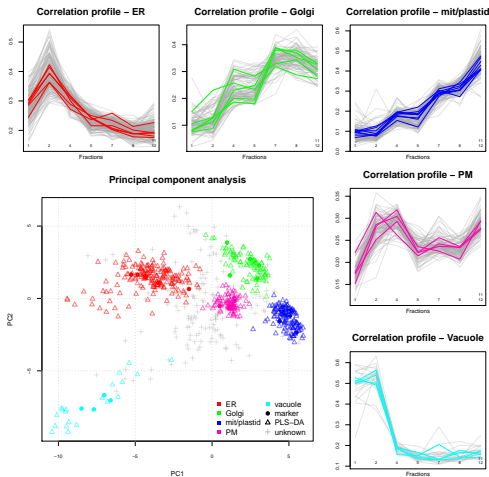|	Fraction$_1$	Fraction$_2$	...	Fraction$_L$	markers
$\mathbf{x}_1$	$x_{1,1}$	$x_{1,2}$	...	$x_{1,L}$	unknown
$\mathbf{x}_2$	$x_{2,1}$	$x_{2,2}$	...	$x_{2,L}$	loc_1
$\mathbf{x}_3$	$x_{3,1}$	$x_{3,2}$	...	$x_{3,L}$	unknown
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{x}_i$	$x_{i,1}$	$x_{i,2}$	...	$x_{i,L}$	loc_k
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{x}_N$	$x_{N,1}$	$x_{N,2}$	...	$x_{N,K}$	unknown

Visualisation

Figure: From Gatto et al. (2010), *Arabidopsis thaliana* data from Dunkley et al. (2006)
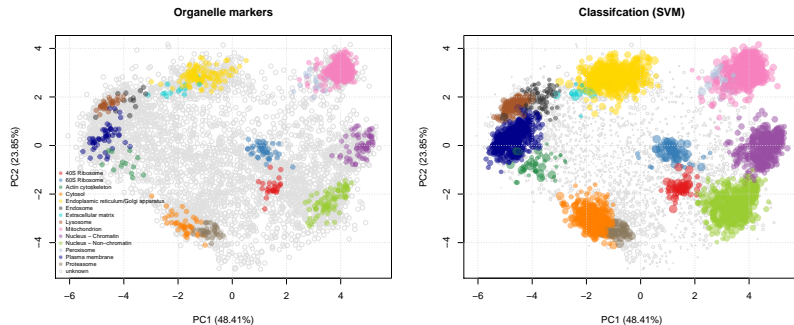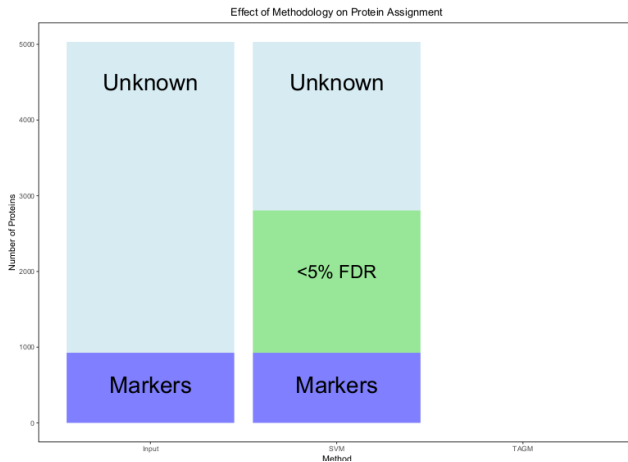
Supervised Machine Learning

Figure: Support vector machines classifier (after 5% FDR classification cutoff) on the embryonic stem cell data from Christoforou et al. (2016).

How much do we learn? How much do we miss?

A Bayesian Mixture Modelling Approach For Spatial Proteomics

- *T Augmented Gaussian Mixture model (TAGM)* is a **multivariate Gaussian generative model** for MS-based spatial proteomics data. It posits that each annotated sub-cellular niche can be modelled by a multivariate Gaussian distribution.

A Bayesian Mixture Modelling Approach For Spatial Proteomics

- *T Augmented Gaussian Mixture model (TAGM)* is a **multivariate Gaussian generative model** for MS-based spatial proteomics data. It 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* (Crook et al., 2018, 2019).

A Bayesian Mixture Modelling Approach For Spatial Proteomics

- *T Augmented Gaussian Mixture model (TAGM)* is a **multivariate Gaussian generative model** for MS-based spatial proteomics data. It 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* (Crook et al., 2018, 2019).

- This methodology allows proteome-wide **uncertainty quantification**, thus adding a further layer to the analysis of spatial proteomics.

We initially model the distribution of profiles associated with proteins that localise to the k-th component as multivariate normal with mean vector $\boldsymbol{\mu}_k$ and covariance matrix Σ_k, so that:

$$\mathbf{x}_i | z_i = k \quad \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k) \tag{1}$$

We initially model the distribution of profiles associated with proteins that localise to the k-th component as multivariate normal with mean vector $\boldsymbol{\mu}_k$ and covariance matrix Σ_k, so that:

$$\mathbf{x}_i | z_i = k \quad \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k) \tag{1}$$

We extend it by introducing an additional *outlier component*. To do this, we augment our model by introducing a further indicator latent variable ϕ. Each protein $\mathbf{x}_i$ is now described by an additional variable ϕ_i, with $\phi_i = 1$ indicating that protein $\mathbf{x}_i$ belongs to a organelle derived component and $\phi_i = 0$ indicating that protein $\mathbf{x}_i$ is not well described by these known components. This outlier component is modelled as a multivariate T distribution with degrees of freedom κ, mean vector $\mathbf{M}$, and scale matrix V.

$$\mathbf{x}_i | z_i = k, \phi_i \quad \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)^{\phi_i} \mathcal{T}(\kappa, \mathbf{M}, V)^{1-\phi_i} \tag{2}$$
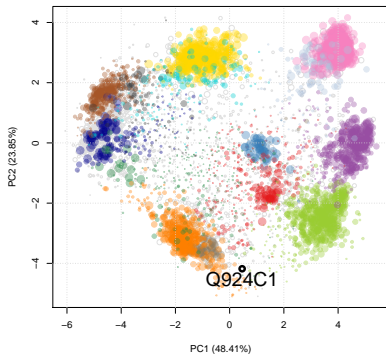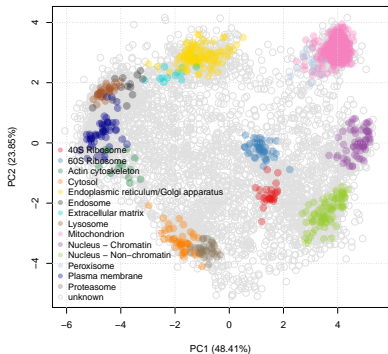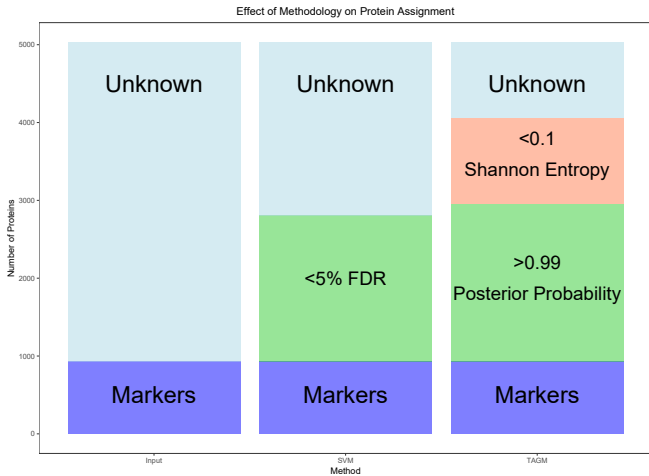
Figure: Assignment of proteins of *unknown* location to one of the annotated classes. The dots are scaled according to the protein assignment probabilities.
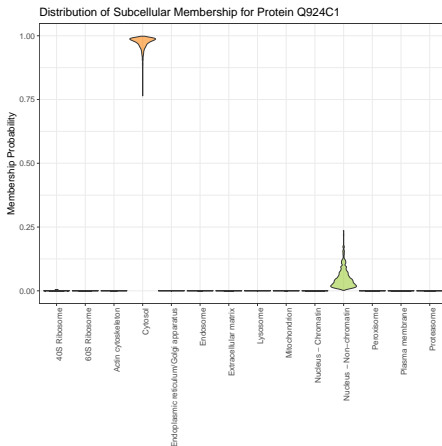
Figure: Exportin 5 (Q924C1) forms part of the micro-RNA export machinery, transporting miRNA from the nucleus to the cytoplasm for further processing. It then translocates back through the nuclear pore complex to return to the nucleus to mediate further transport between nucleus and cytoplasm. The model correctly 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 reflects the multi-location of the protein.

Whole sub-cellular proteome uncertainty
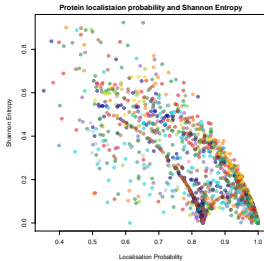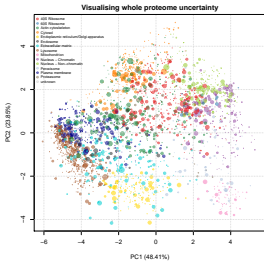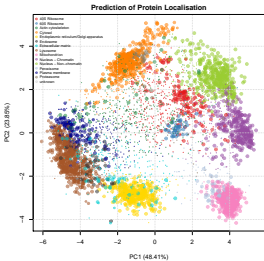
Conclusions

- Protein sub-cellular localisation: *localisation is function*.

- Reliance on computational biology, statistics and dedicated software (for example `MSnbase` (Gatto and Lilley, 2012), `pRoloc` (Gatto et al., 2014)) to interpret data and acquire biological knowledge (details not shown).

- Rigorous computational infrastructure and sound data analysis and interpretation is a **long term investment**.

References I

A Christoforou, C M Mulvey, L M Breckels, A Geladaki, T Hurrell, P C Hayward, T Naake, L Gatto, R Viner, A Martinez Arias, and K S Lilley. A draft map of the mouse pluripotent stem cell spatial proteome. *Nat Commun*, 7:8992, Jan 2016. doi: 10.1038/ncomms9992.

Oliver M. Crook, Claire M. Mulvey, Paul D. W. Kirk, Kathryn S. Lilley, and Laurent Gatto. A bayesian mixture modelling approach for spatial proteomics. *PLOS Computational Biology*, 14(11):1–29, 11 2018. doi: 10.1371/journal.pcbi.1006516. URL https://doi.org/10.1371/journal.pcbi.1006516.

OM Crook, LM Breckels, KS Lilley, PDW Kirk, and L Gatto. A bioconductor workflow for the bayesian analysis of spatial proteomics [version 1; peer review: 1 approved, 2 approved with reservations]. *F1000Research*, 8(446), 2019. doi: 10.12688/f1000research.18636.1.

TPJ Dunkley, S Hester, IP Shadforth, J Runions, T Weimar, SL Hanton, JL Griffin, C Bessant, F Brandizzi, C Hawes, RB Watson, P Dupree, and KS Lilley. Mapping the Arabidopsis organelle proteome. *PNAS*, 103(17):6518–6523, Apr 2006.

LJ Foster, CL de Hoog, Y Zhang, Y Zhang, X Xie, VK Mootha, and M Mann. A mammalian organelle map by protein correlation profiling. *Cell*, 125(1):187–199, Apr 2006.

L Gatto and KS Lilley. MSnbase - an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. *Bioinformatics*, 28(2):288–9, Jan 2012.

References II

L Gatto, JA Vizcaino, H Hermjakob, W Huber, and KS Lilley. Organelle proteomics experimental designs and analysis. *Proteomics*, 2010.

L Gatto, L M Breckels, S Wieczorek, T Burger, and K S Lilley. Mass-spectrometry based spatial proteomics data analysis using pRoloc and pRolocdata. *Bioinformatics*, Jan 2014.

Aikaterini Geladaki, Nina Kocevar Britovsek, Lisa M. Breckels, Tom S. Smith, Claire M. Mulvey, Oliver M. Crook, Laurent Gatto, and Kathryn S. Lilley. LOPIT-DC: A simpler approach to high-resolution spatial proteomics. *bioRxiv*, 2018. doi: 10.1101/378364. URL https://www.biorxiv.org/content/early/2018/07/26/378364.

D N Itzhak, S Tyanova, J Cox, and G H Borner. Global, quantitative and dynamic mapping of protein subcellular localization. *Elife*, 5, Jun 2016. doi: 10.7554/eLife.16950.

C M Mulvey, L M Breckels, A Geladaki, N K Britovek, DJH Nightingale, A Christoforou, M Elzek, M J Deery, L Gatto, and K S Lilley. Using hyperlopit to perform high-resolution mapping of the spatial proteome. *Nat Protoc*, 12(6): 1110–1135, Jun 2017. doi: 10.1038/nprot.2017.026.

Acknowledgements

- **Mr Oliver Crook** and **Dr Lisa Breckels**, (U of Cambridge): spatial proteomics, machine learning, software.
- Funding: BBSRC (UK), Wellcome Trust (UK)

Thank you for your attention