Probabilistic modelling of protein sub-cellular localisation
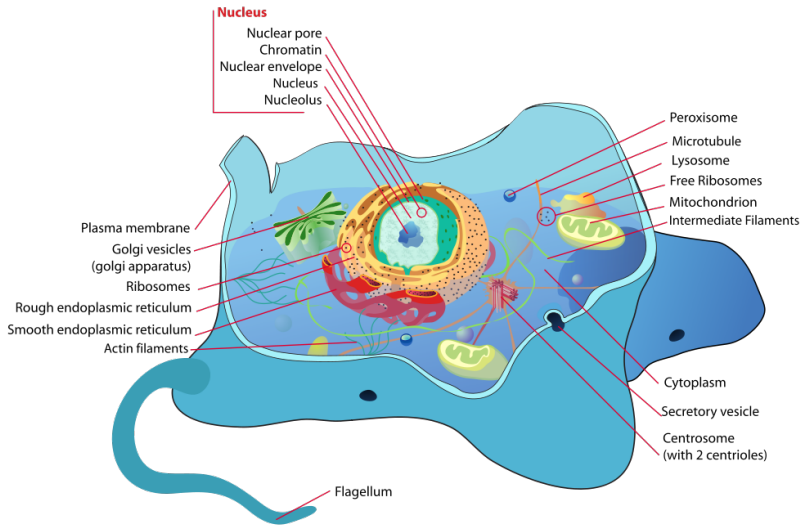
Laurent Gatto

laurent.gatto@uclouvain.be

http://lgatto.github.io/about

de Duve Institute – UCLouvain

13 June 2019 – The Alan Turing Institute
Workshop on Statistical Data Science for
Proteomics and Metabolomics

Use case: spatial proteomics.

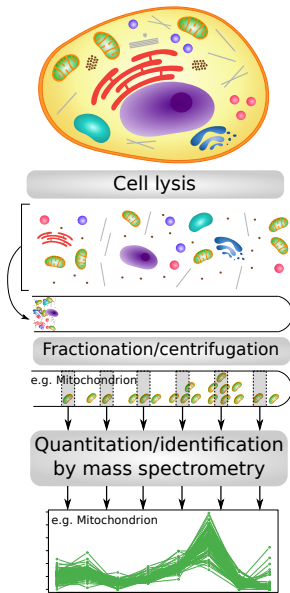# Cell organisation - regulation of protein localisation

Spatial proteomics is the systematic study of protein localisations.

Experiment

Gradient approaches: PCP
(Dunkley et al., 2006), LOPIT
(Foster et al., 2006),
hyperLOPIT (Christoforou et al.,
2016; Mulvey et al., 2017).

**Explorative/discovery
approaches**, steady-state **global
localisation maps**.

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

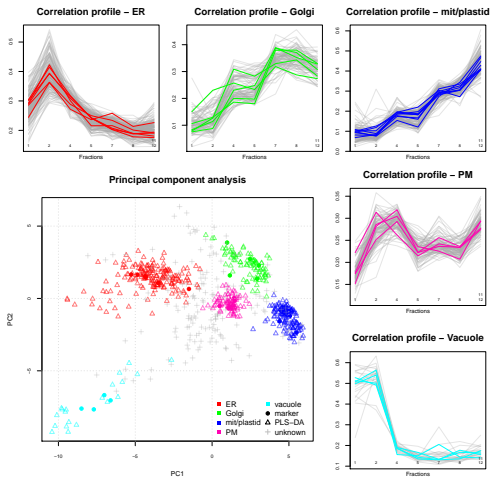|	Fraction$_1$	Fraction$_2$	...	Fraction$_L$	markers
$\mathbf{x}_1$	$x_{1,1}$	$x_{1,2}$	...	$x_{1,L}$	unknown
$\mathbf{x}_2$	$x_{2,1}$	$x_{2,2}$	...	$x_{2,L}$	loc_1
$\mathbf{x}_3$	$x_{3,1}$	$x_{3,2}$	...	$x_{3,L}$	unknown
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{x}_i$	$x_{i,1}$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

Data analysis

Visualisation

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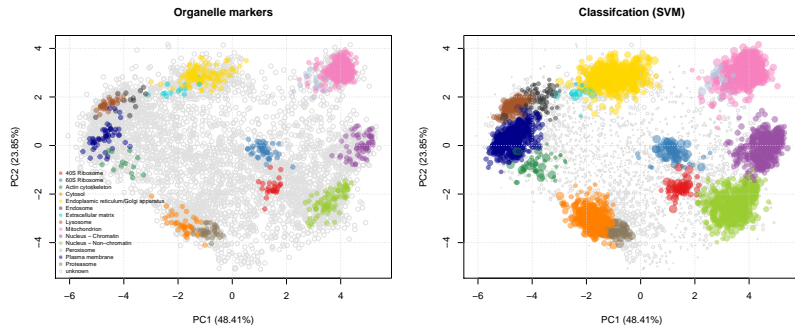
Supervised Machine Learning

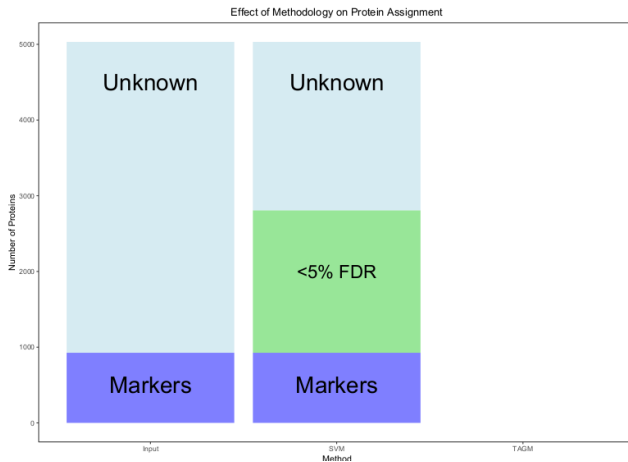

Figure: Support vector machines classifier (after 5% FDR classification cutoff) on the embryonic stem cell data from Christoforou et al. (2016).

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 (Crook et al., 2018, 2019). 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 (Crook et al., 2018, 2019). 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*.

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 (Crook et al., 2018, 2019). 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*.

- 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.

$$\mathbf{x}_i|z_i = k, \phi_i \quad \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)^{\phi_i} \mathcal{T}(\kappa, \mathbf{M}, V)^{1-\phi_i} \tag{2}$$
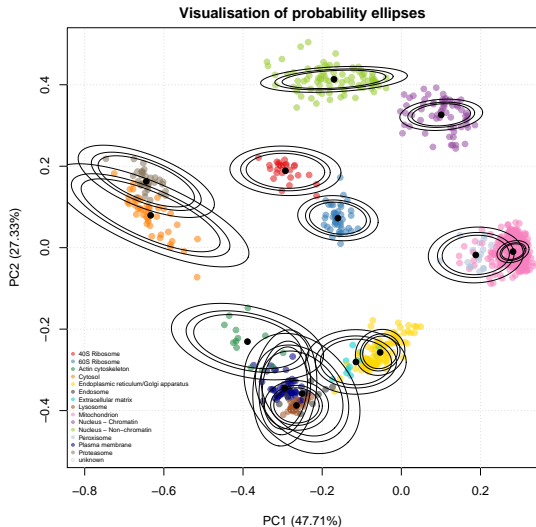
Visualisation of probability ellipses

Figure: 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.
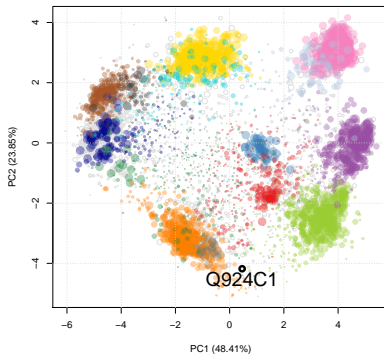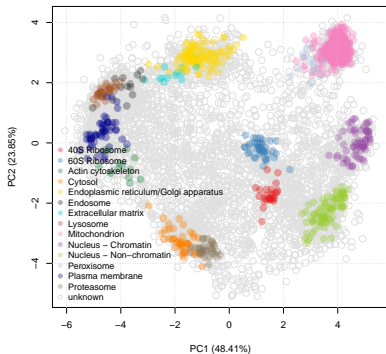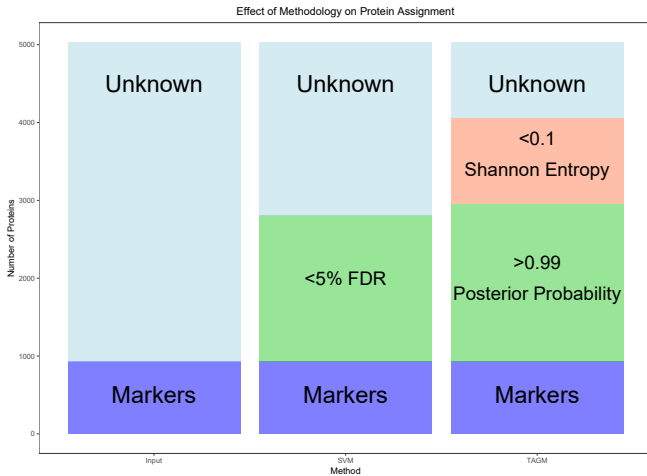
Figure: Assignment of proteins of *unknown* location to one of the annotated classes. The dots are scaled according to the protein assignment probabilities.
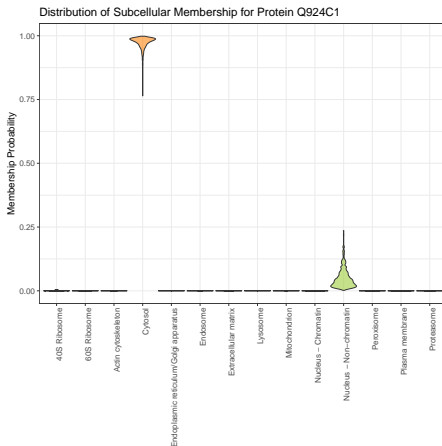
Effect of Methodology on Protein Assignment

Distribution of Subcellular Membership for Protein Q924C1

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
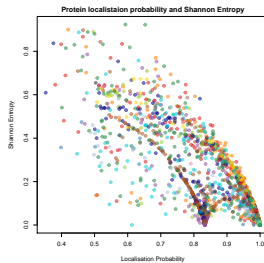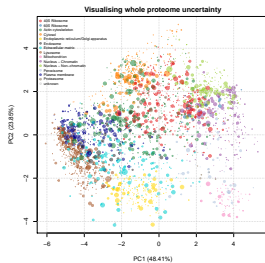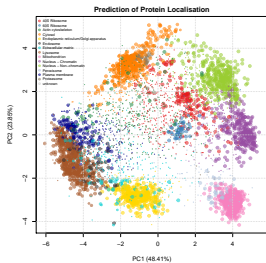
Behind the scenes: software/data structures and open research practice.

Beyond the figures[1]

- Software: **infrastructure** (MSnbase, Gatto and Lilley (2012)), **dedicated machine learning** (pRoloc, Gatto et al. (2014)), **interactive visualisation**[2] (pRolocGUI, Breckels et al. (2017)) and **data** (pRolocdata, Gatto et al. (2014)) for spatial proteomics.

- **Open source**, and **coordinated and collaborative development**, enabling **reproducible research**, enables understanding of the data (not a black box) and **drive scientific innovation**.

---

[1] ... which are all reproducible, by the way.
[2] https://lgatto.shinyapps.io/christoforou2015/

Conclusions

- Protein sub-cellular localisation: technologies (hyperLOPIT) and opportunities.

- Reliance on computational biology, statistics and dedicated software (pRoloc *et al.*) to interpret data and acquire biological knowledge.

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

References I

Lisa Breckels, Thomas Naake, and Laurent Gatto. *pRolocGUI: Interactive visualisation of spatial proteomics data*, 2017. URL http://ComputationalProteomicsUnit.github.io/pRolocGUI/. R package version 1.11.2.

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.

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

L Gatto, LM Breckels, T Burger, DJ Nightingale, AJ Groen, C Campbell, N Nikolovski, CM Mulvey, A Christoforou, M Ferro, and KS Lilley. A foundation for reliable spatial proteomics data analysis. *MCP*, 13(8):1937–52, Aug 2014.

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.
- **Dr Sebastian Gibb** and **Dr Johannes Rainer**: MS and proteomics software.
- Prof Kathryn Lilley (U of Cambridge), Dr Claire Mulvey, (CRUK Cambridge Institute): data.
- Funding: BBSRC, Wellcome Trust.
- Slides available under CC-BY SA license at https://github.com/lgatto/2019_06_13_StatDS_Turing

Thank you for your attention