Mapping the sub-cellular proteome

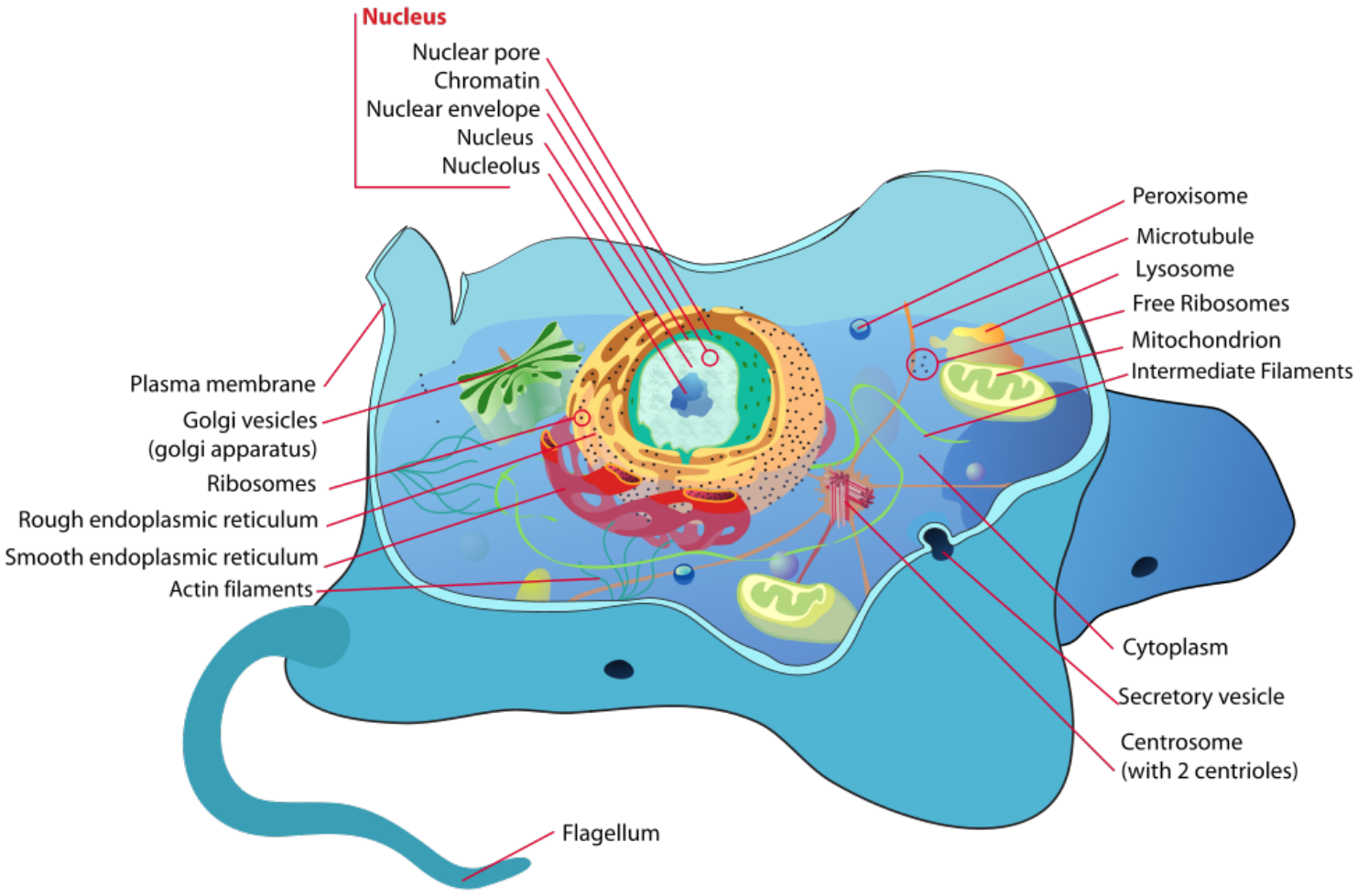
Probabilistic modelling of protein sub-cellular localisation

Laurent Gatto
laurent.gatto@uclouvain.be
http://lgatto.github.io/about
de Duve Institute – UCLouvain

Slides: http://bit.ly/20190830pfs (CC-BY)

Protein Folding and Stability
30 August 2019 – Liège

Cell organisation - **localisation is function**



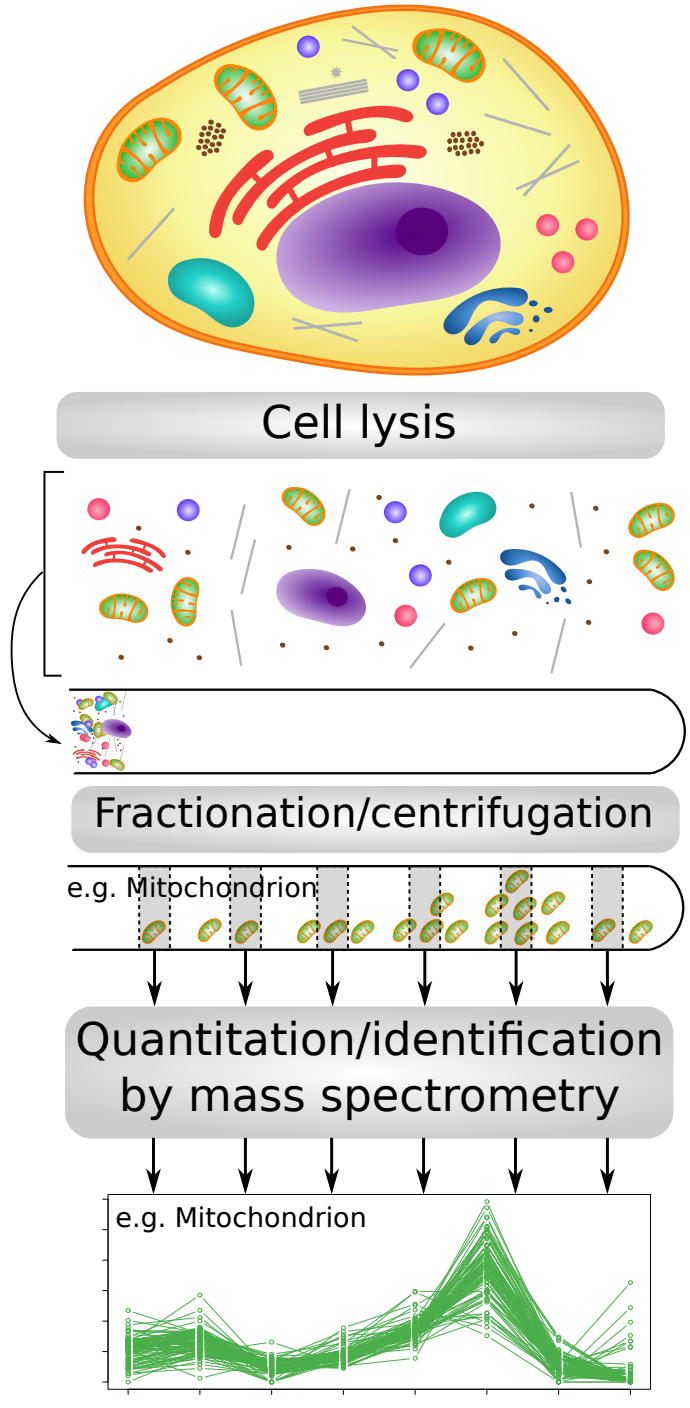
Spatial proteomics is the systematic study of protein localisations.

Localisation(s) – re-localisation – mis-localisation

Image from Wikipedia http://en.wikipedia.org/wiki/Cell_(biology).

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).



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

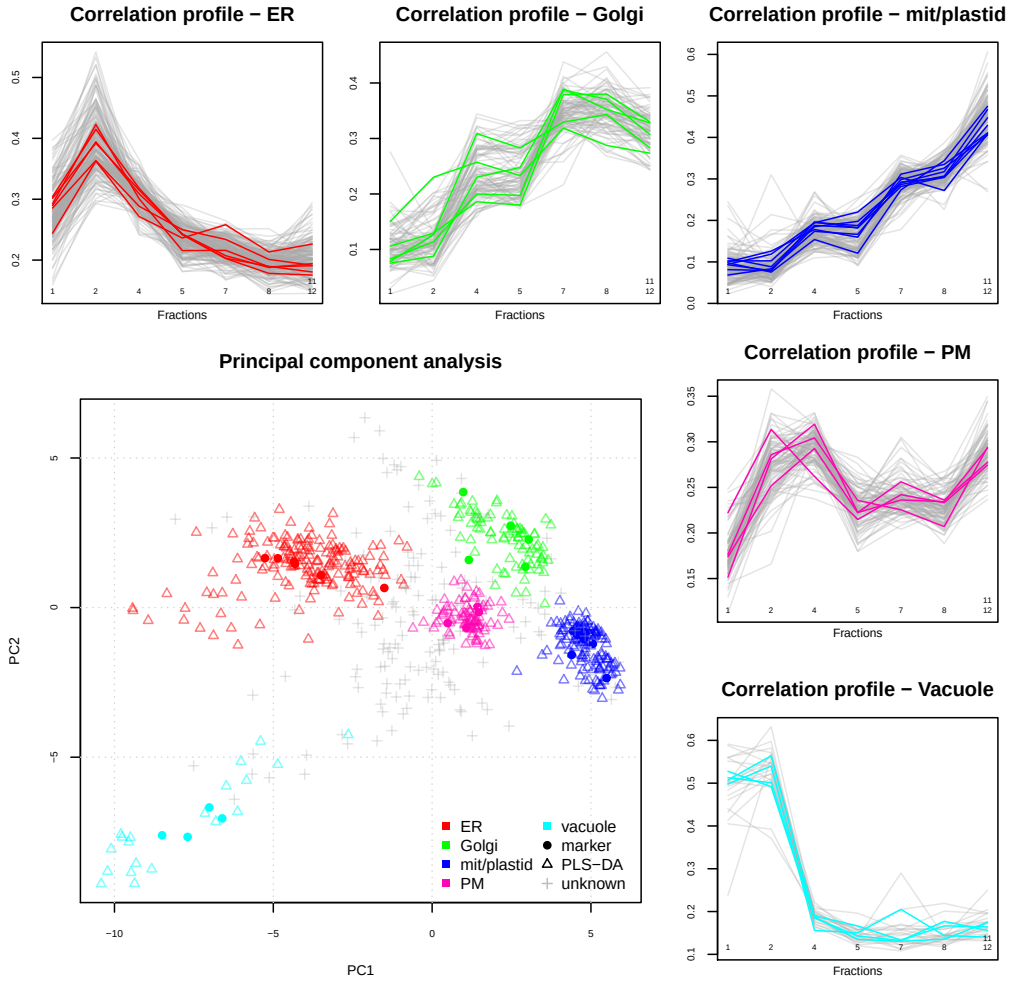


Figure: From Gatto et al. (2010), *Arabidopsis thaliana* data from Dunkley et al. (2006)

Supervised Machine Learning to infer **localisation**

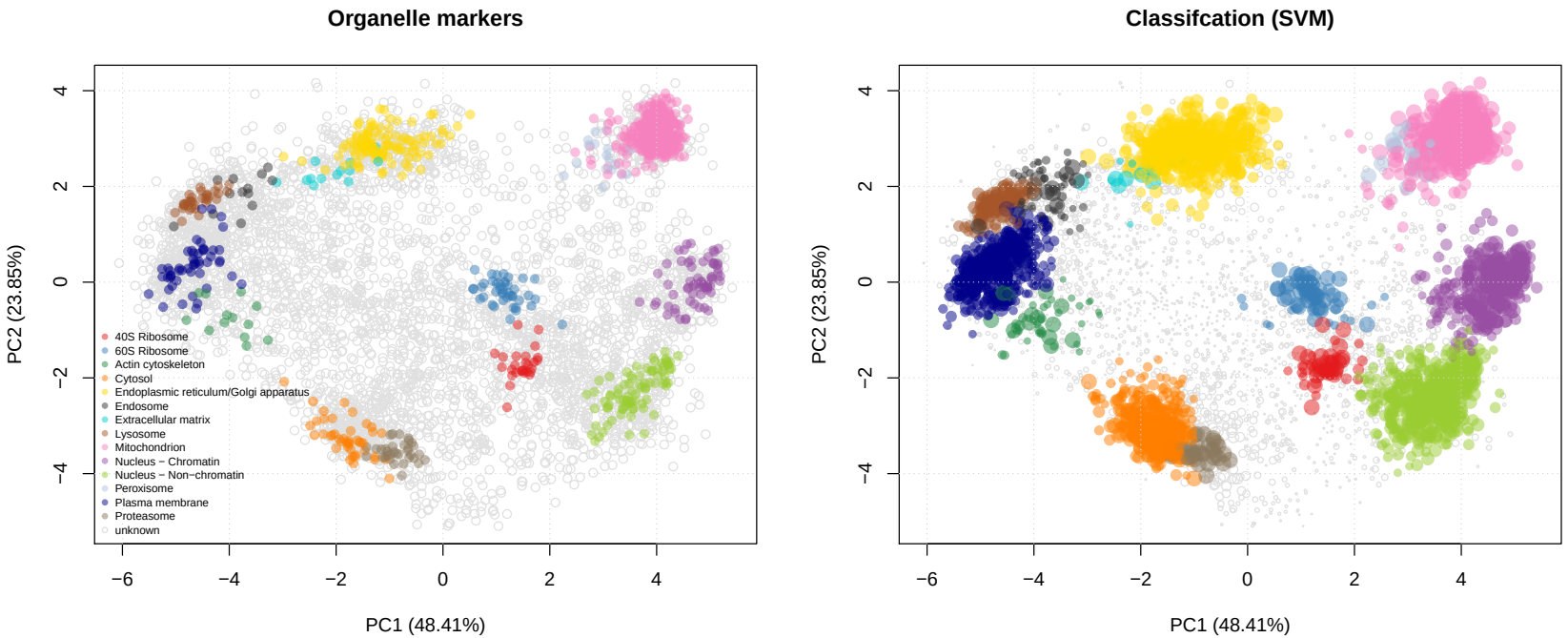
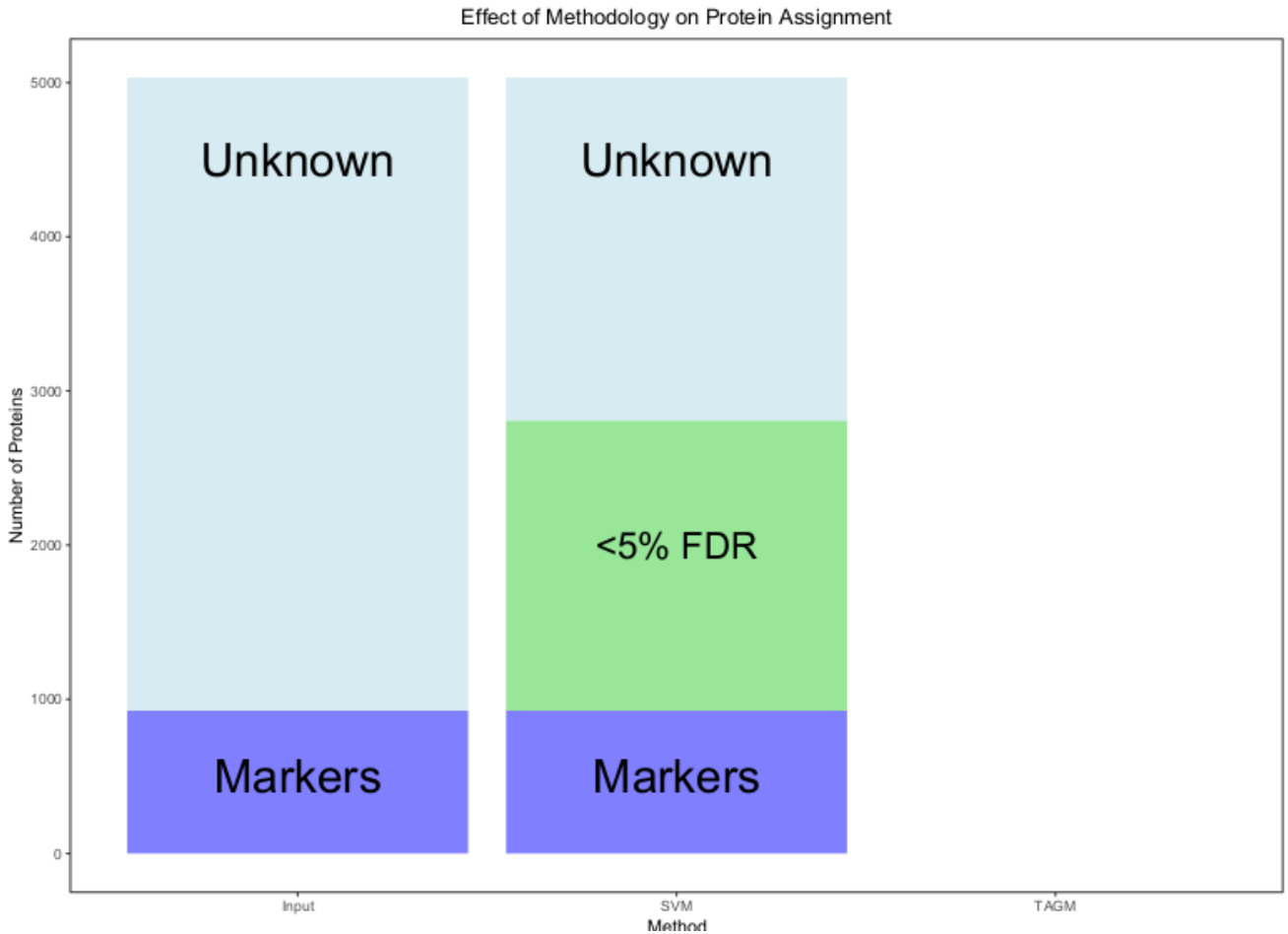


Figure: Support vector machines classifier (after 5% FDR classification cutoff) on the embryonic stem cell data from Christoforou et al. (2016).

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.

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

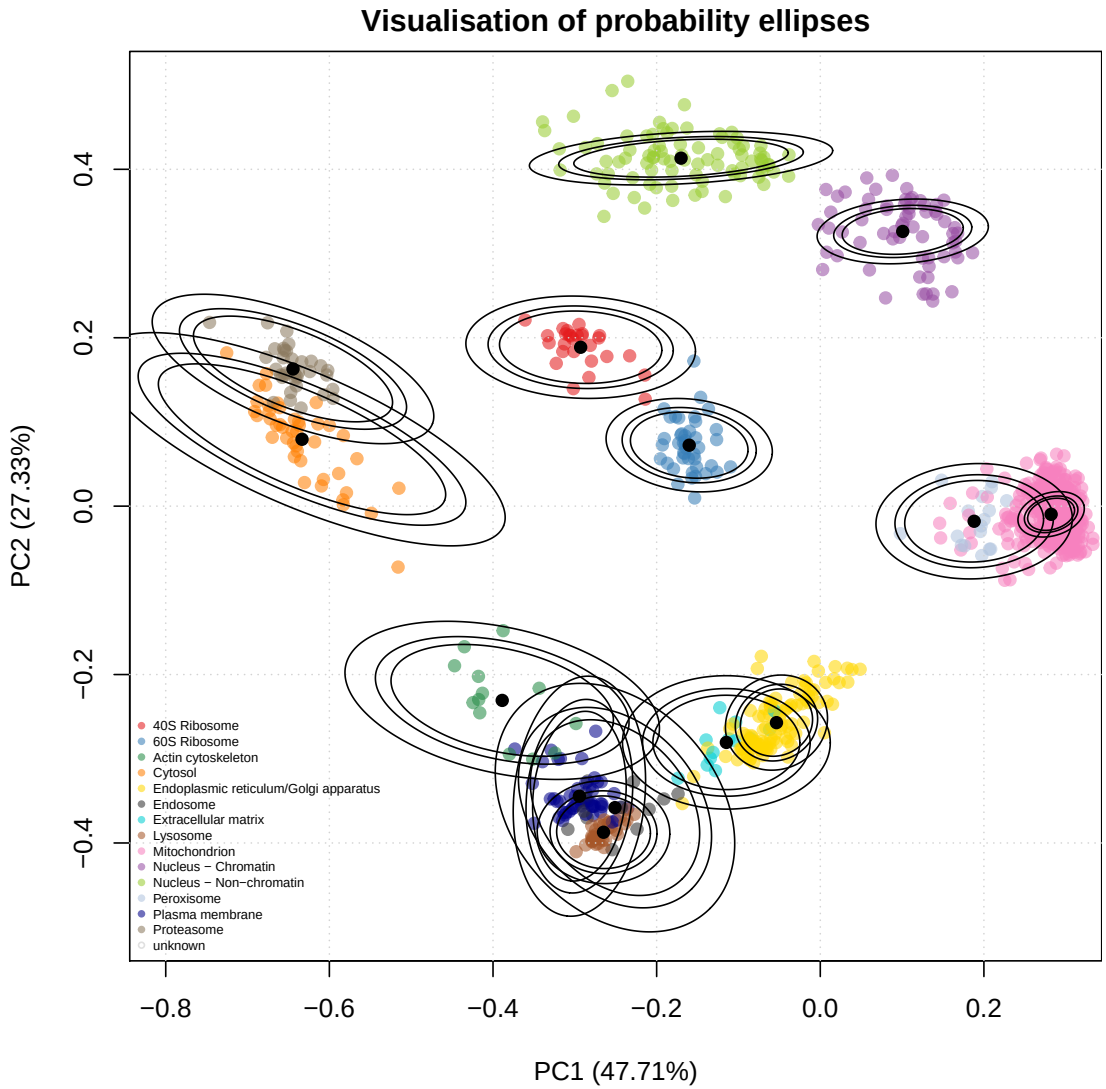


Figure: Illustration of how the TAGM model describes the pluripotent mouse embryonic stem cell data. Each ellipse contains a proportion of total probability of a particular multivariate Gaussian density. The outer ellipse contains 99% of the total probability whilst the middle and inner ellipses contain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

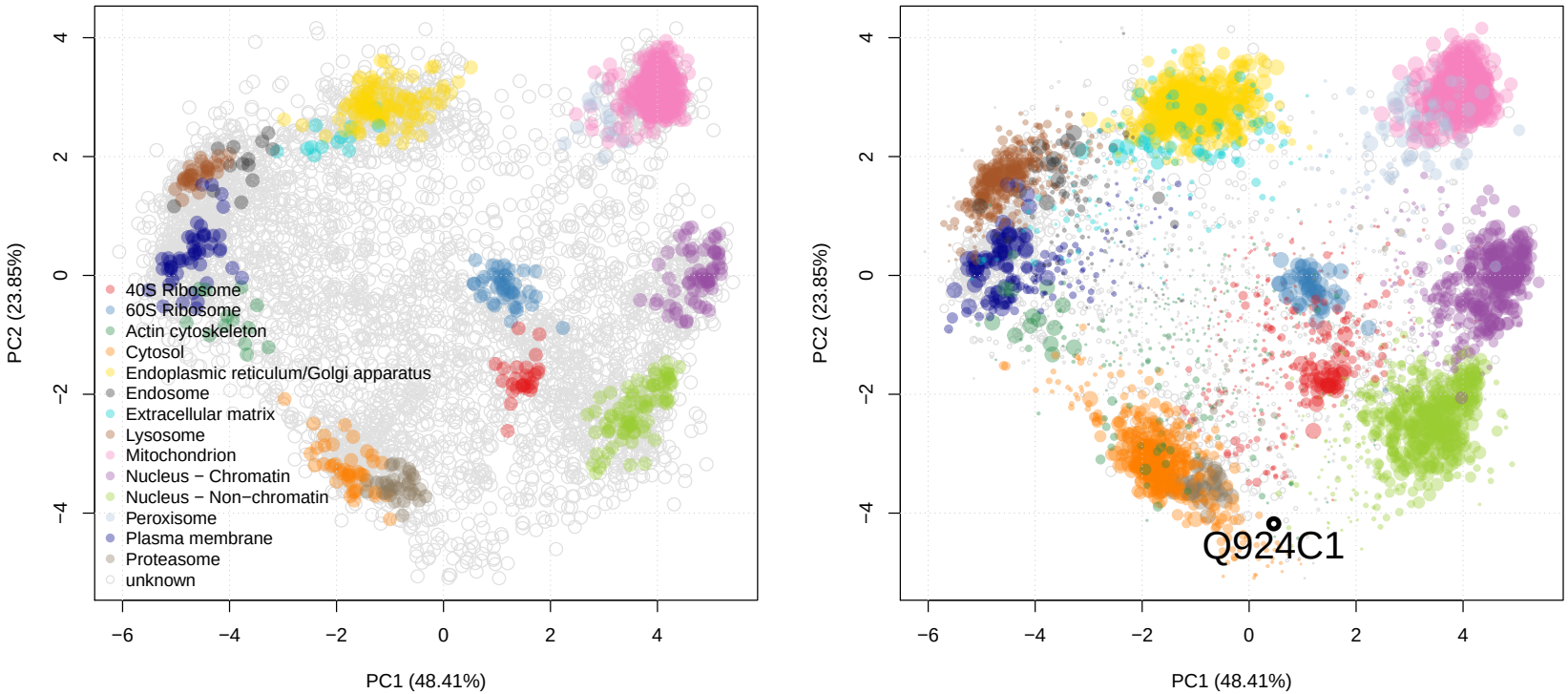
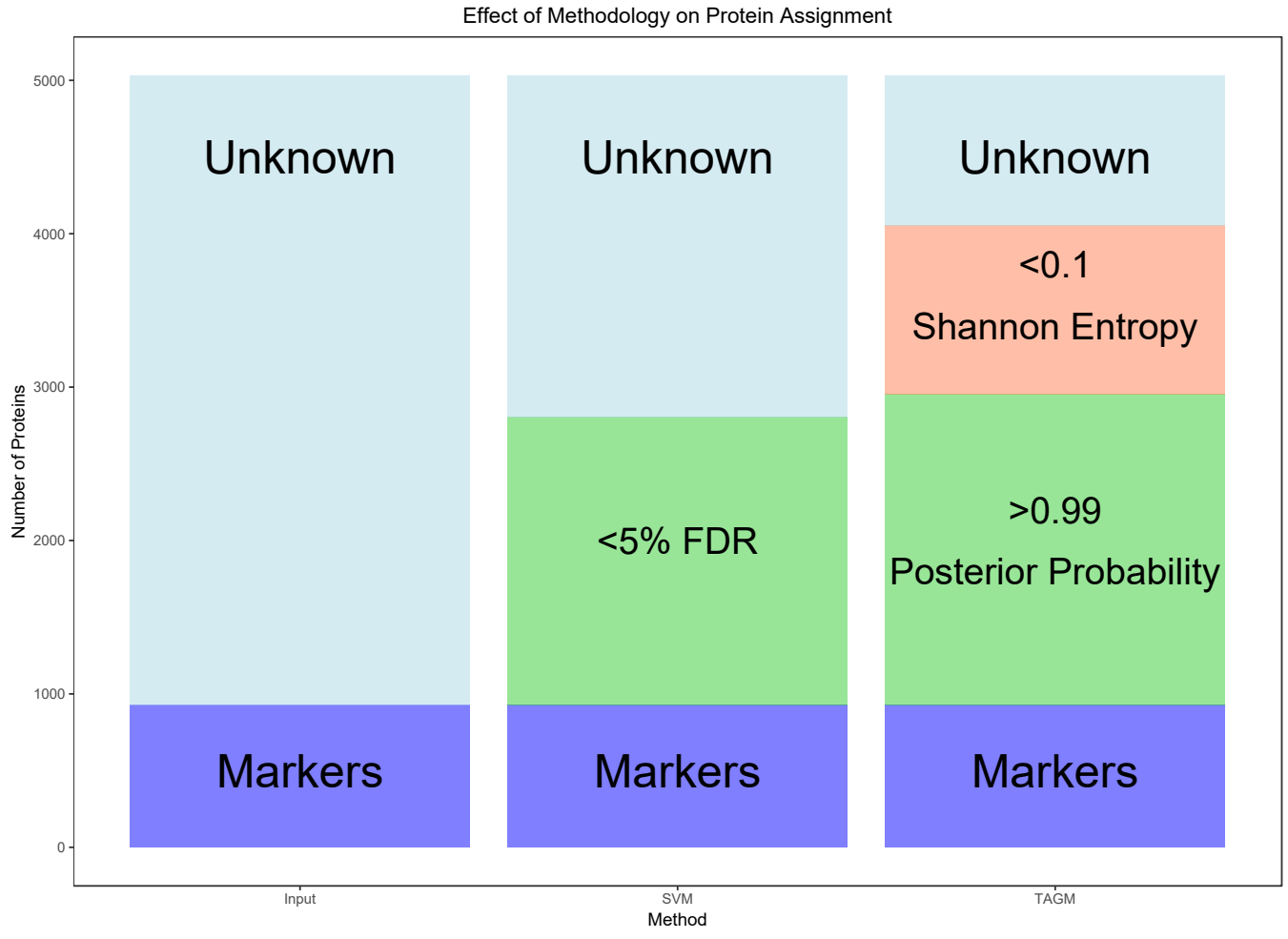


Figure: Assignment of proteins of *unknown* location to one of the annotated classes. The dots are scaled according to the protein assignment probabilities.


Effect of Methodology on Protein Assignment
Number of Proteins
5000
4000
3000
2000
1000
0
Unknown
Unknown
Unknown
<0.1
Shannon Entropy
<5% FDR
>0.99
Posterior Probability
Markers
Markers
Markers
Input
SVM
TAGM
Method

Multi-localisation: **localisations**

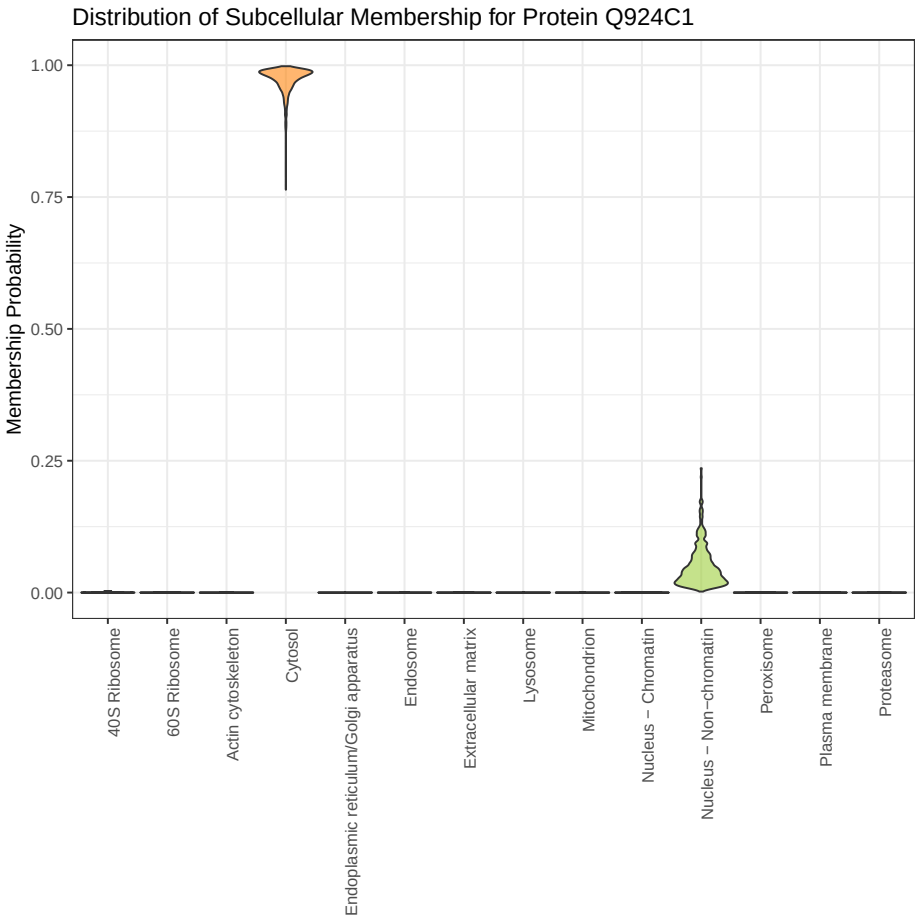
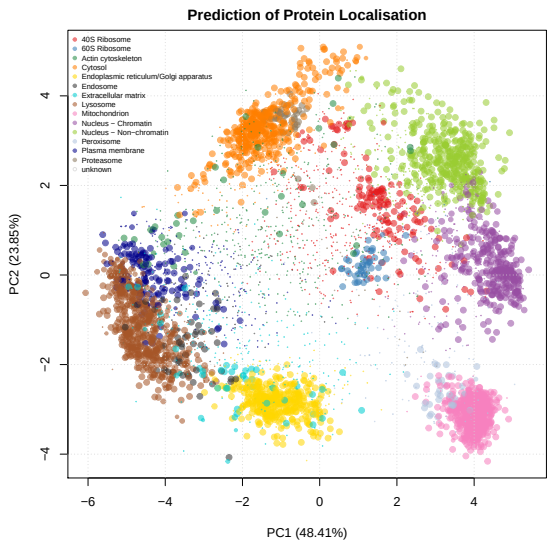
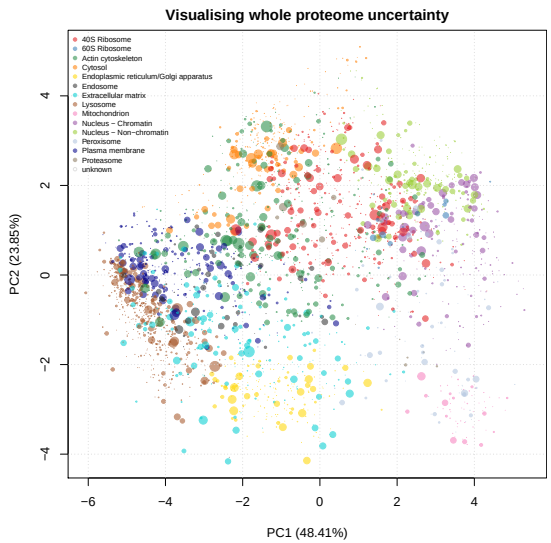
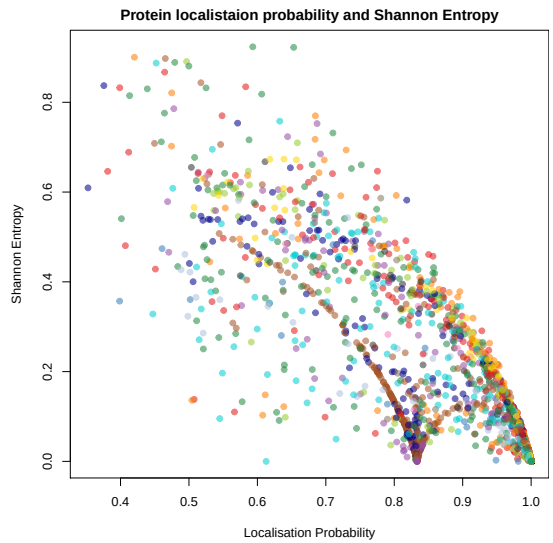


Figure: Exportin 5 (Q924C1) forms part of the micro-RNA export machinery, transporting miRNA from the nucleus to the cytoplasm for further processing. It then translocates back through the nuclear pore complex to return to the nucleus to mediate further transport between nucleus and cytoplasm. The model 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







Spatial dynamics

Trans-localisation event during monocyte to macrophage differentiation

Investigate the effect of lipopolysaccharides (LPS)-mediated inflammatory response in human monocytic cells (THP-1)

Data

- Triplicate **temporal** profiling (0, 2, 4, 6, 12, 24 hours).
- Triplicate **spatial** profiling (0 vs 12 hours) - early trafficking, before actual morphological differentiation at 24h.

Work lead by **Dr Claire Mulvey**, Cambridge.

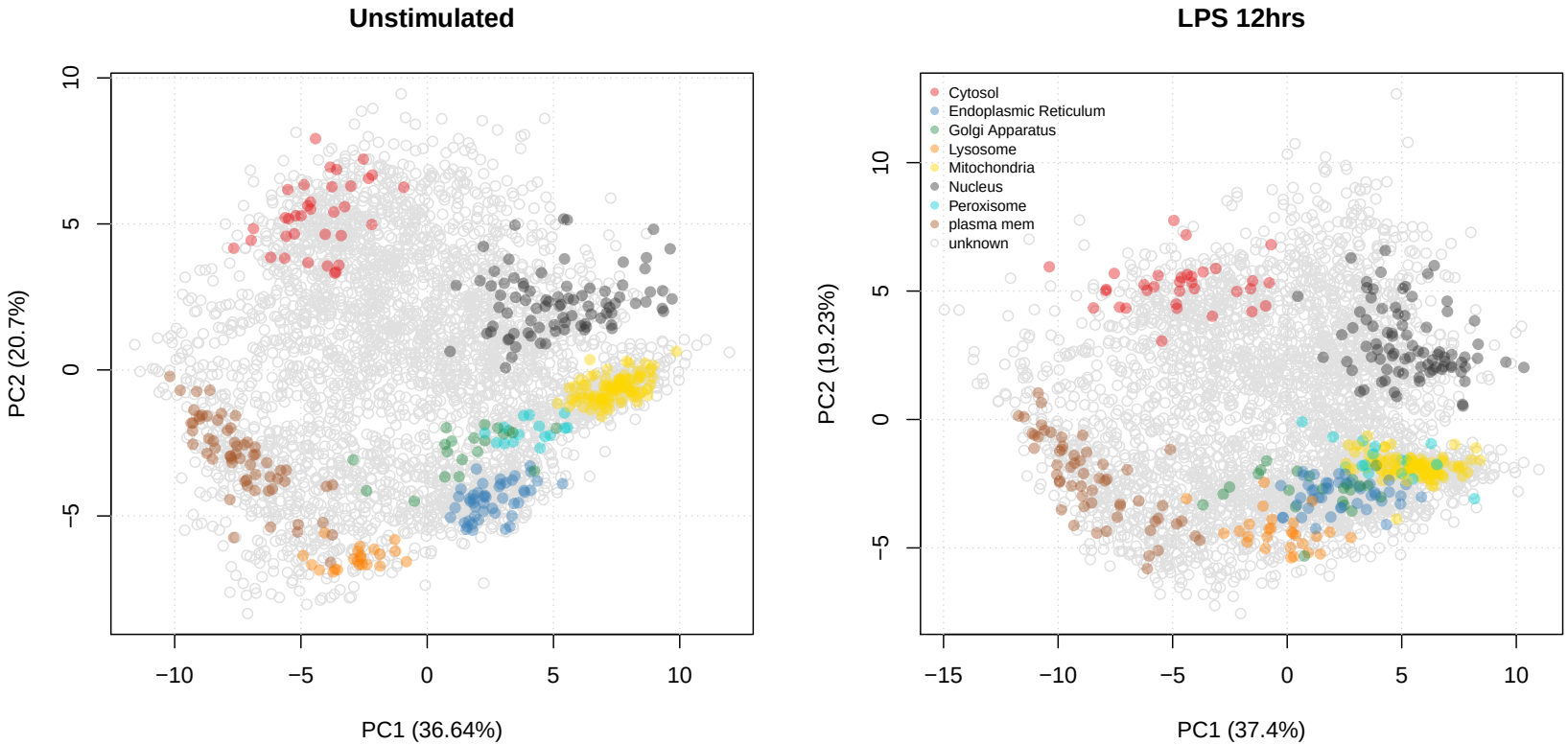


Figure: Spatial maps of unstimulated and LPS-treated cells (combined triplicates).

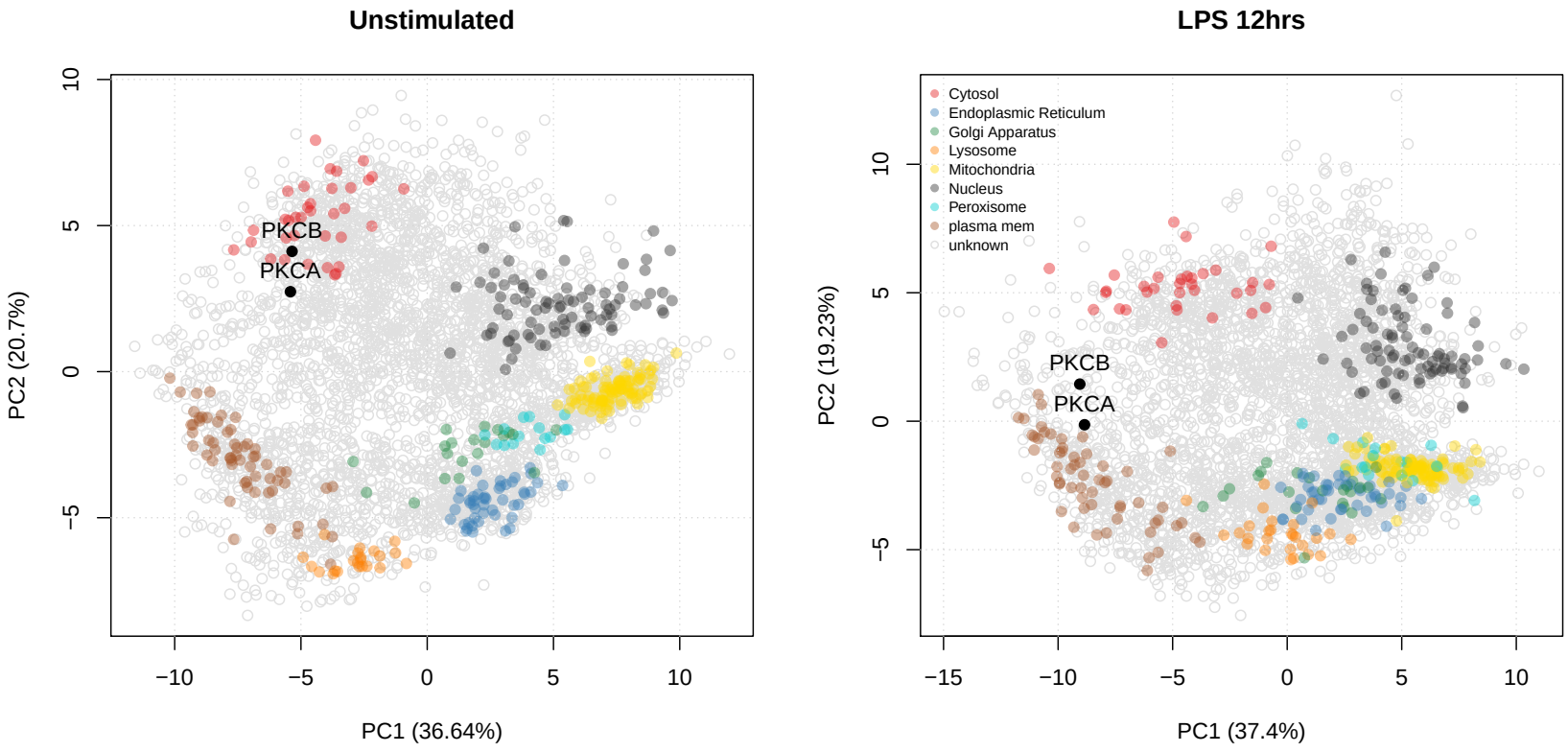


Figure: Relocation of Protein Kinase C α and β from the cytosol to the plasma membrane, **driving maturation into a differentiated macrophage phenotype**.

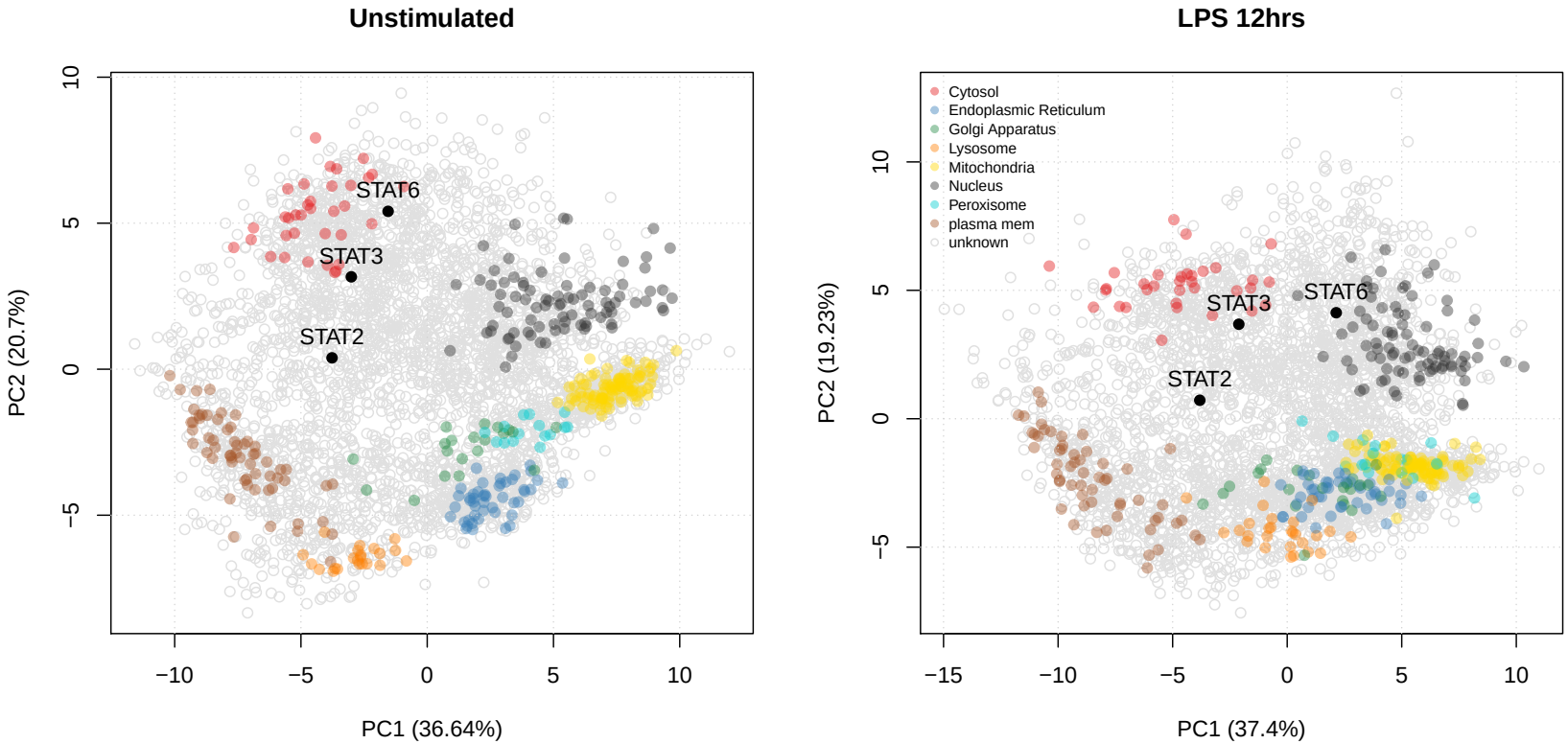


Figure: Relocation of Signal transducer and activator of transcription 6 (STAT6) from the cytosol to the Nucleus, **activating anti-bacterial and anti-viral-like response**. Validated by microscopy and see also Chen et al. (2011).

Folding and stability

- Re-localisation upon protein post-translational modification.
- Different sub-cellular micro-environments driving different conformations.
- Effect of localisation on protein structure.

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. (2014b)), **interactive visualisation**[2] (pRolocGUI, Breckels et al. (2017)) and **data** (pRolocdata, Gatto et al. (2014b)) for spatial proteomics.

[1] ... which are all reproducible, by the way.

[2] https://lgatto.shinyapps.io/christoforou2015/

[3] between and within domains/software

Beyond the figures[1]

- Software: **infrastructure** (MSnbase, Gatto and Lilley (2012)), **dedicated machine learning** (pRoloc, Gatto et al. (2014b)), **interactive visualisation**[2] (pRolocGUI, Breckels et al. (2017)) and **data** (pRolocdata, Gatto et al. (2014b)) for spatial proteomics.
- The **Bioconductor** (Huber et al., 2015) ecosystem for high throughput biology data analysis and comprehension: **open source**, and **coordinated and collaborative**[3] **open 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/

[3] between and within domains/software

Open research: open source software

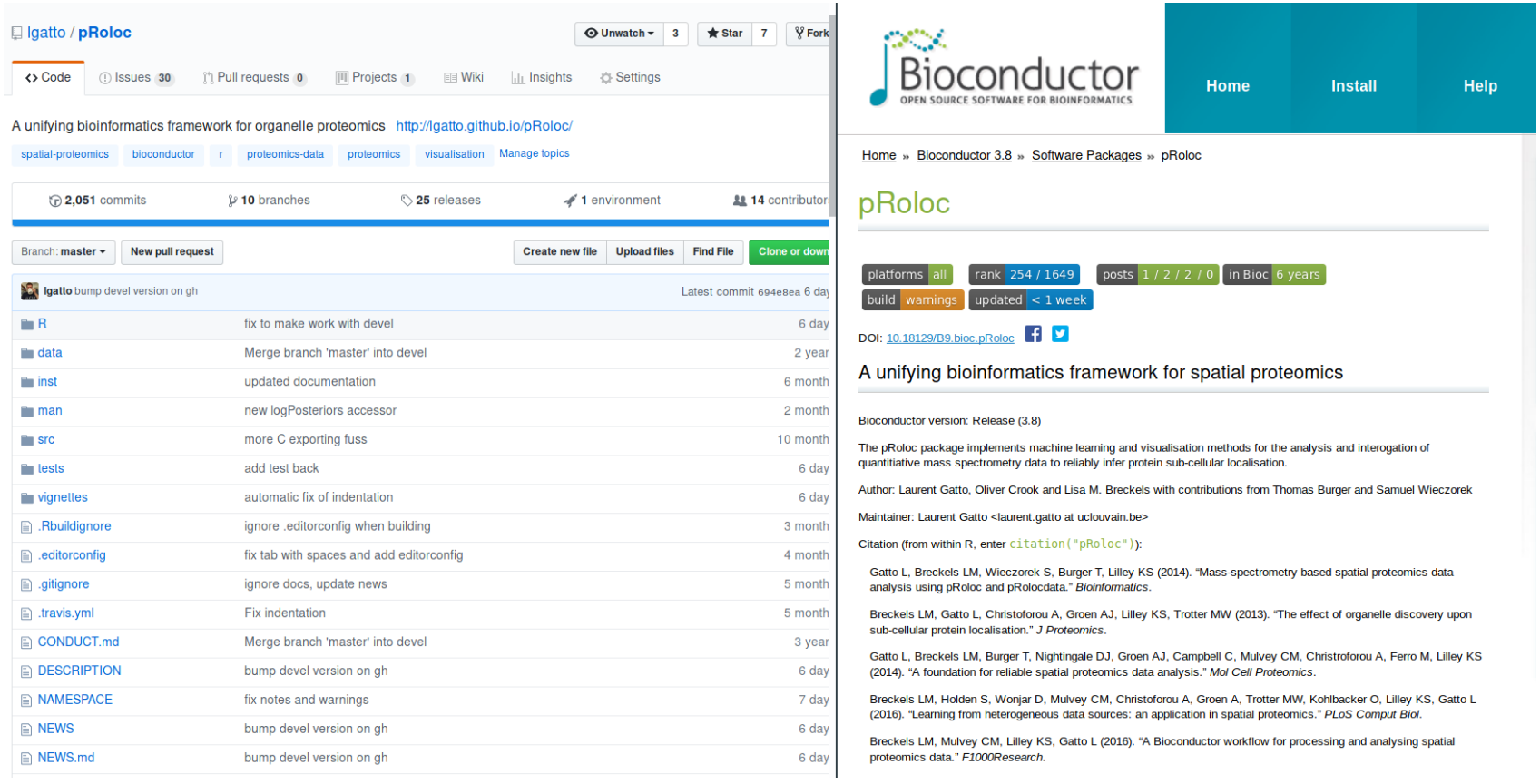


Figure: Gatto et al. (2014a) Left: Public repository for the `pRoloc` software (`https://github.com/lgatto/pRoloc`). Right: offical Bioconductor page.

Open and reproducible research

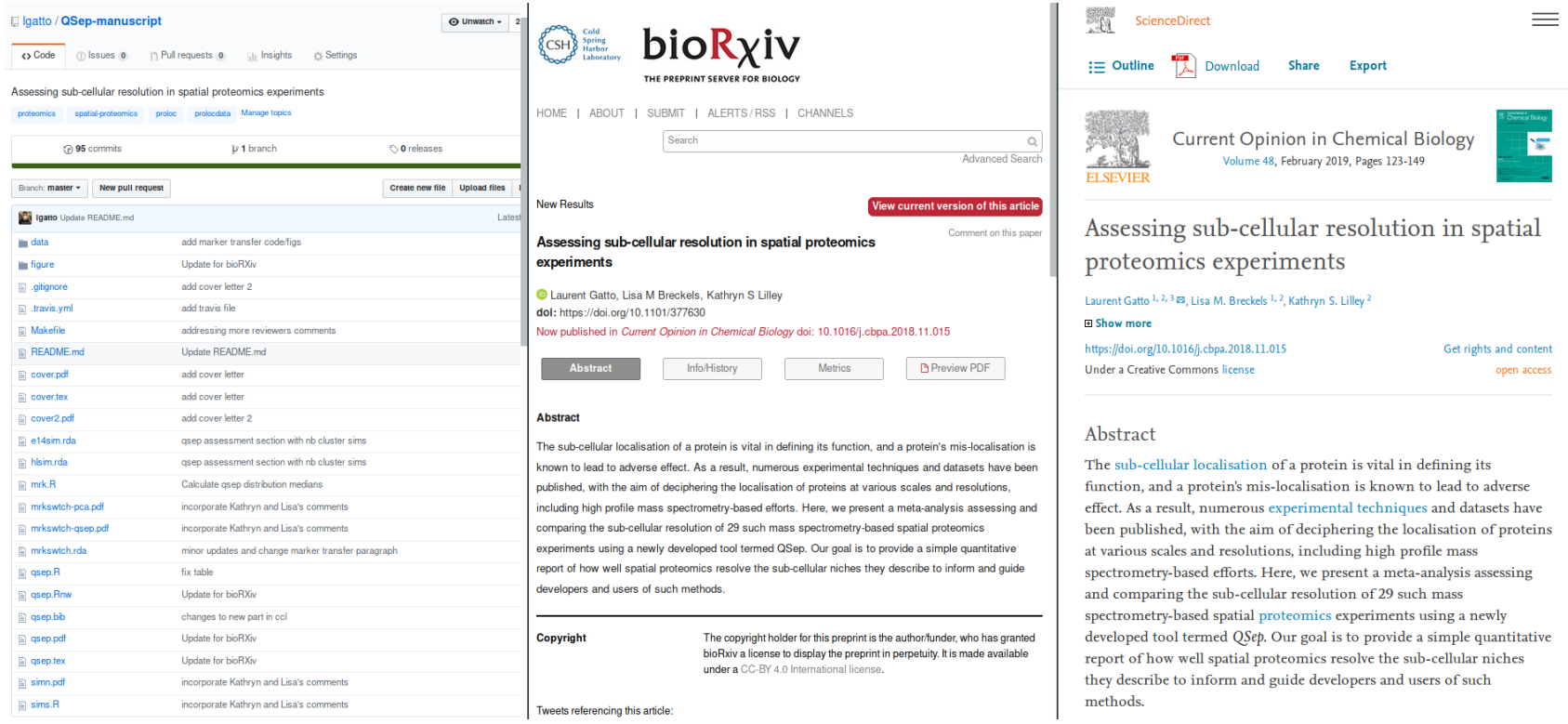


Figure: Gatto et al. (2018) reproducible document (https://github.com/lgatto/QSep-manuscript), preprint (https://doi.org/10.1101/377630) and paper (https://doi.org/10.1016/j.cbpa.2018.11.015).

Working with open and reproducible research in mind doesn't mean releasing everything prematurely, it means

- managing research in a way one can find data and results at every stage
- one can reproduce results, re-run/compare them with new data or different methods/parameters, and
- one can release data (or parts thereof) when/if appropriate.

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., 2014a)) 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

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.

H Chen, H Sun, F You, W Sun, X Zhou, L Chen, J Yang, Y Wang, H Tang, Y Guan, W Xia, J Gu, H Ishikawa, D Gutman, G Barber, Z Qin, and Z Jiang. Activation of stat6 by sting is critical for antiviral innate immunity. *Cell*, 147(2):436–46, Oct 2011. doi: 10.1016/j.cell.2011.09.022.

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.

References II

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

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 2014b.

Laurent Gatto, Lisa M Breckels, and Kathryn S Lilley. Assessing sub-cellular resolution in spatial proteomics experiments. *bioRxiv*, 2018. doi: 10.1101/377630.

References III

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.

W Huber, V J Carey, R Gentleman, S Anders, M Carlson, B S Carvalho, H C Bravo, S Davis, L Gatto, T Girke, R Gottardo, F Hahne, K D Hansen, R A Irizarry, M Lawrence, M I Love, J MacDonald, V Obenchain, A K Oleś, H Pagès, A Reyes, P Shannon, G K Smyth, D Tenenbaum, L Waldron, and M Morgan. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods*, 12(2):115–21, Jan 2015. doi: 10.1038/nmeth.3252.

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.
- Kathryn Lilley (U of Cambridge): spatial proteomics data.
- Funding: BBSRC (UK), Wellcome Trust (UK), FNRS (BE)

Slides: `http://bit.ly/20190830pfs` (CC-BY)

Thank you for your attention