Probabilistic modelling of protein sub-cellular localisation

Laurent Gatto
de Duve Institute – UCLouvain
laurent.gatto@uclouvain.be – @lgatt0
http://lgatto.github.io/about

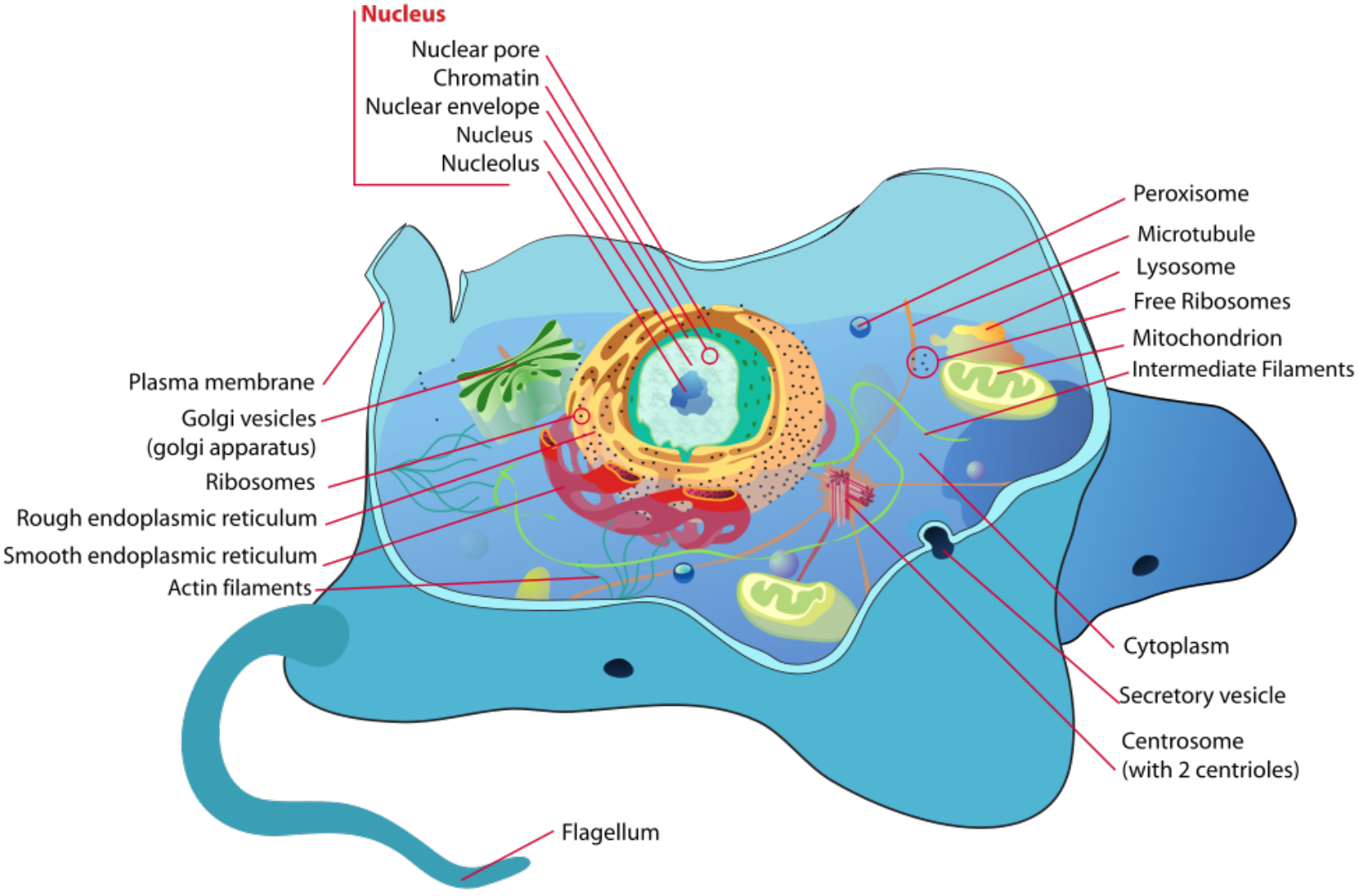
Slides: DOI 10.5281/zenodo.1435058
https://zenodo.org/record/1435058

26 September 2018, Gent

Abstract

In biology, localisation is function - understanding the sub-cellular localisation of proteins is paramount to comprehend the context of their functions. Mass spectrometry-based spatial proteomics and contemporary machine learning enable to build proteome-wide spatial maps, 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 us 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.

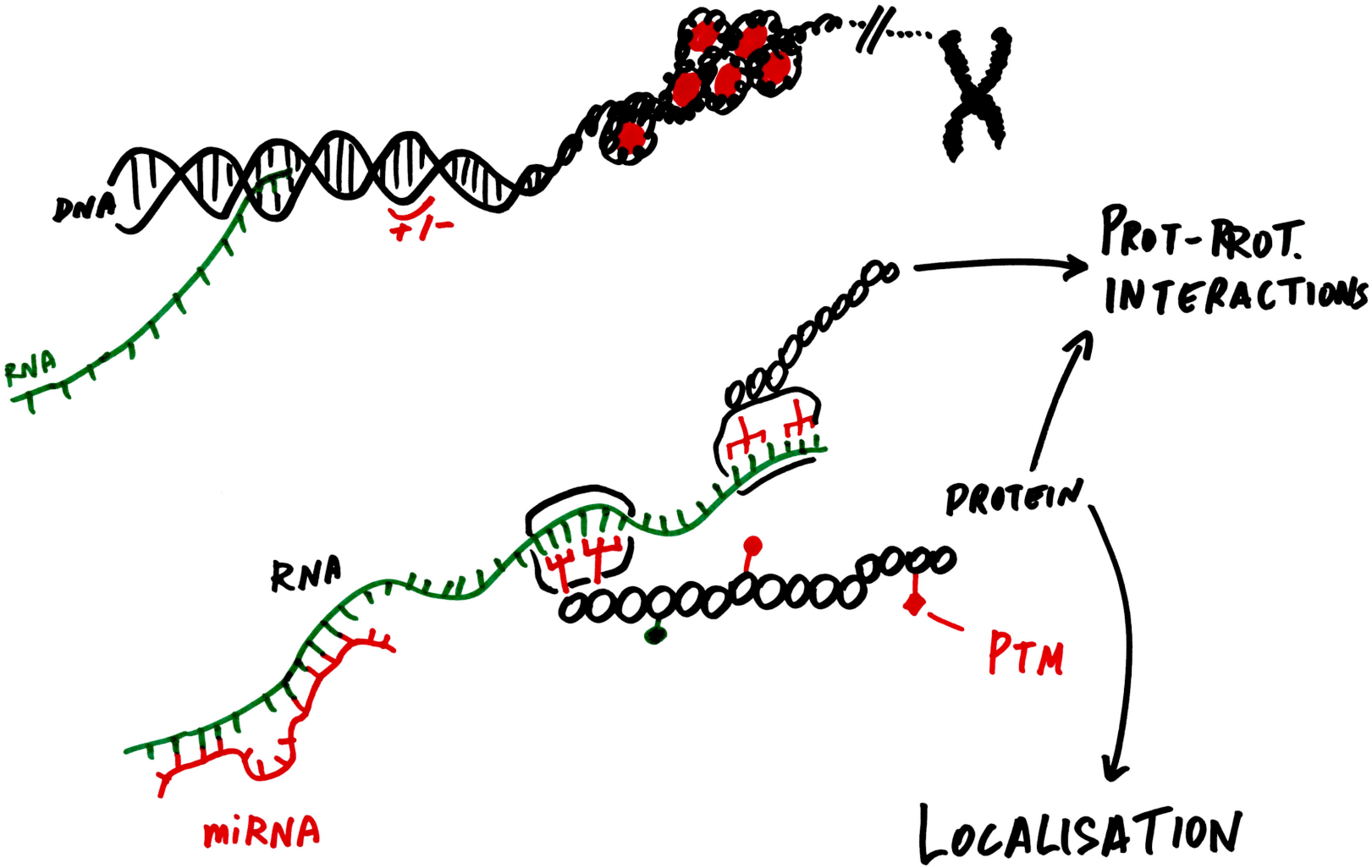
Cell organisation



Spatial proteomics is the systematic study of protein localisations.

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

Regulations



Spatial proteomics - Why?

Localisation is function

- The cellular sub-division allows cells to establish a range of distinct micro-environments, each favouring different biochemical reactions and interactions and, therefore, allowing each compartment to fulfil a particular functional role.
- Localisation and sequestration of proteins within sub-cellular niches is a fundamental mechanism for the post-translational regulation of protein function.

Re-localisation in

- Differentiation stem cells.
- Activation of biological processes.

Spatial proteomics - Why?

Mis-localisation

Disruption of the targeting/trafficking process alters proper sub-cellular localisation, which in turn perturb the cellular functions of the proteins.

- Abnormal protein localisation leading to the **loss of functional** effects in diseases (Laurila and Vihinen, 2009).
- Disruption of the nuclear/cytoplasmic transport (nuclear pores) have been detected in many types of **carcinoma cells** (Kau et al., 2004).
- Sub-cellular localisation of MC4R with ADCY3 at neuronal primary cilia underlies a common pathway for genetic predisposition to **obesity** (Siljee et al., 2018).

Spatial proteomics - How, experimentally

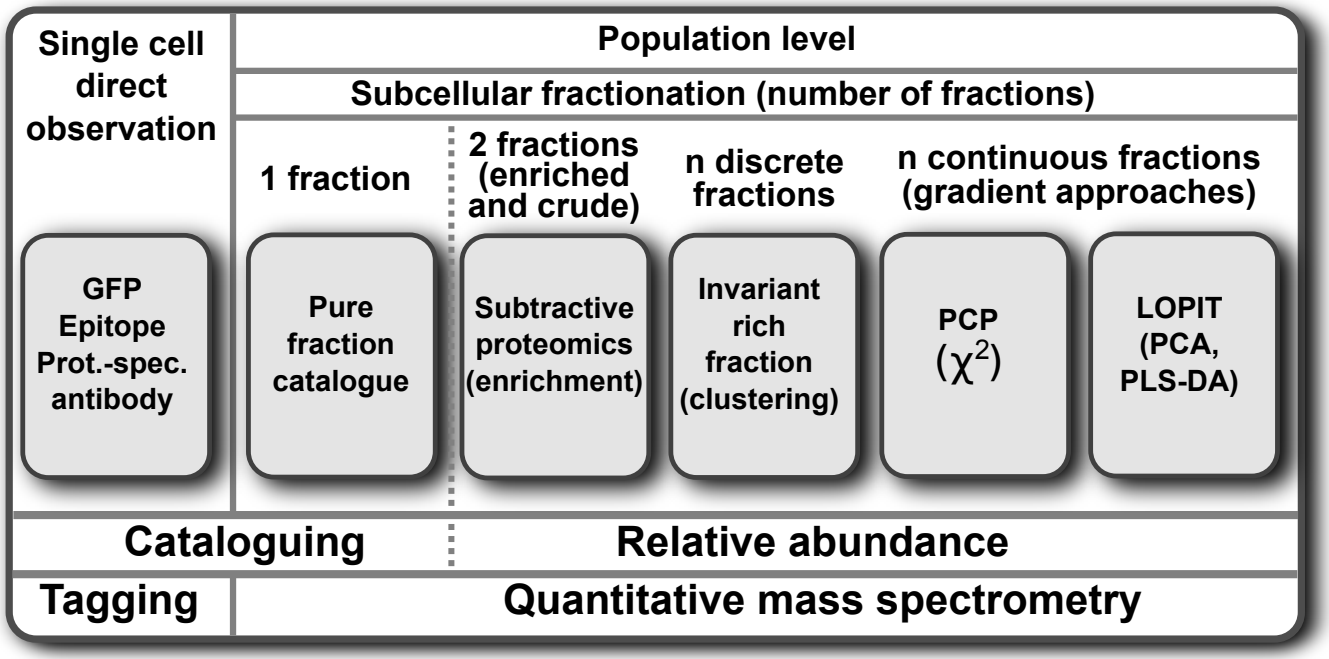


Figure : Organelle proteomics approaches (Gatto et al., 2010)

Fusion proteins and immunofluorescence

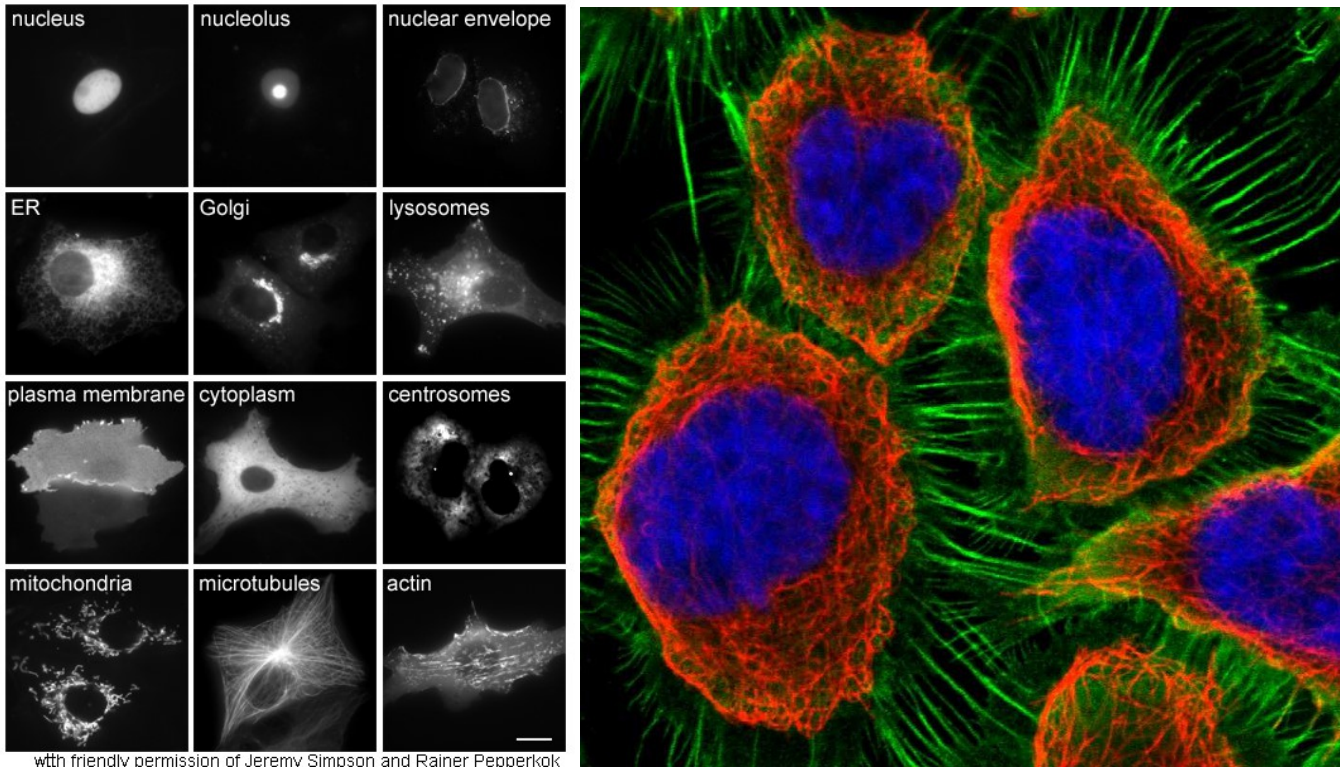


Figure : Targeted protein localisation. Example of discrepancies between IF and FPs as well as between FP tagging at the N and C termini (Stadler et al., 2013).

Spatial proteomics - How, experimentally

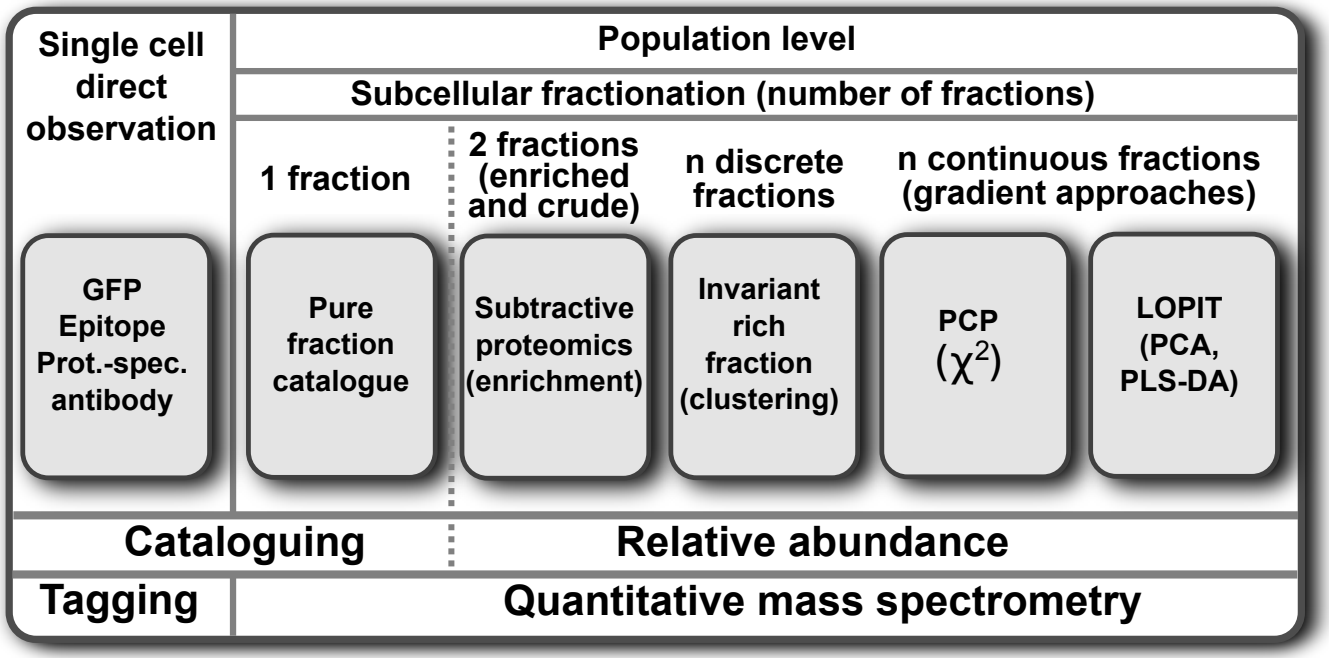
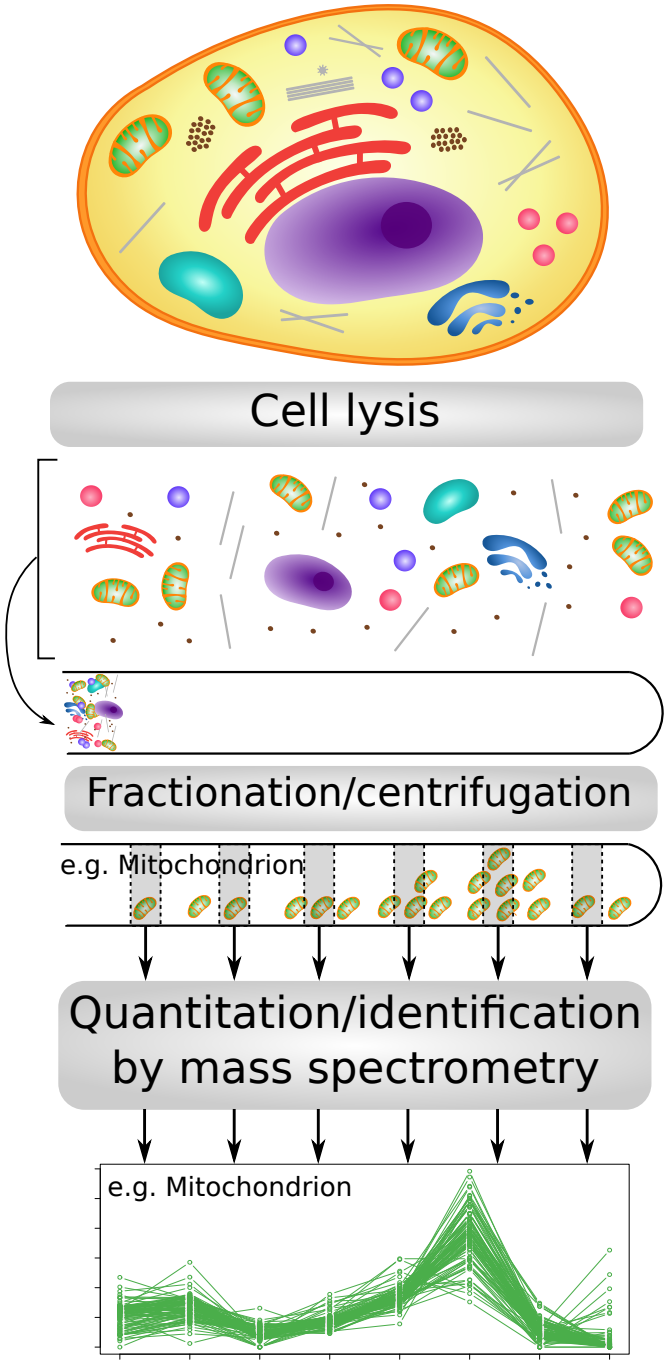


Figure : Organelle proteomics approaches (Gatto et al., 2010).

Gradient approaches: Dunkley et al. (2006), Foster et al. (2006), based on works by de Duve, Claude and Palade.

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


Cell lysis
Fractionation/centrifugation
e.g. Mitochondrion
Quantitation/identification
by mass spectrometry
e.g. Mitochondrion

Quantitation data

	Fraction_1	Fraction_2	...	Fraction_m
p_1	$q_{1,1}$	$q_{1,2}$	...	$q_{1,m}$
p_2	$q_{2,1}$	$q_{2,2}$	...	$q_{2,m}$
p_3	$q_{3,1}$	$q_{3,2}$	...	$q_{3,m}$
p_4	$q_{4,1}$	$q_{4,2}$	...	$q_{4,m}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
p_j	$q_{j,1}$	$q_{j,2}$	...	$q_{j,\ m}$

Quantitation data and organelle markers

	Fraction_1	Fraction_2	...	Fraction_m	markers
p_1	$q_{1,1}$	$q_{1,2}$	...	$q_{1,m}$	unknown
p_2	$q_{2,1}$	$q_{2,2}$	...	$q_{2,m}$	loc_1
p_3	$q_{3,1}$	$q_{3,2}$	...	$q_{3,m}$	unknown
p_4	$q_{4,1}$	$q_{4,2}$	...	$q_{4,m}$	loc_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
p_j	$q_{j,1}$	$q_{j,2}$	...	$q_{j,\ m}$	unknown

Data analysis

- Visualisation (unsupervised learning, clustering) (Gatto et al., 2018)
- Classification (supervised learning) (Breckels et al., 2016b)
- Novelty detection (semi-supervised learning) (Breckels et al., 2013)
- Data integration (transfer learning) (Breckels et al., 2016a)
- Probabilistic modelling (Crook et al., 2018)

Visualisation

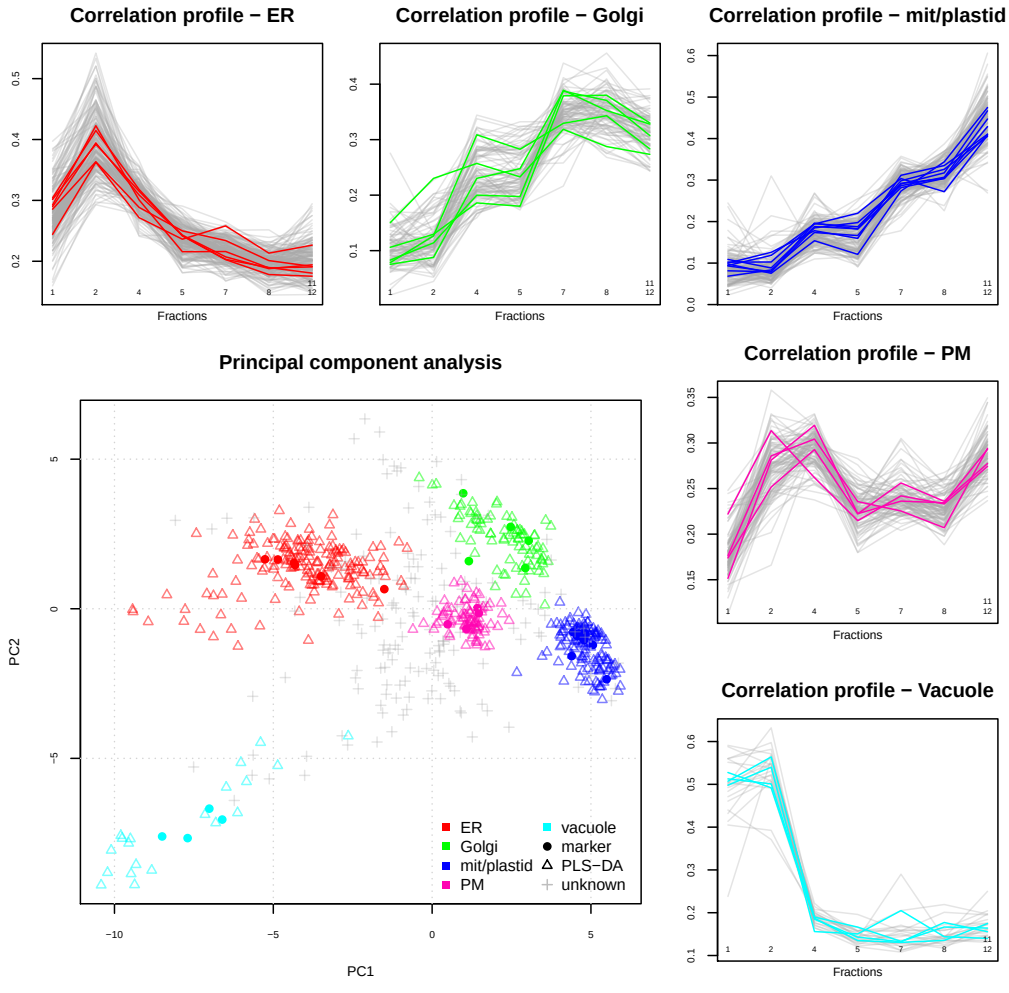


Figure : From Gatto et al. (2010), *Arabidopsis thaliana* data from Dunkley et al. (2006)

Supervised Machine Learning

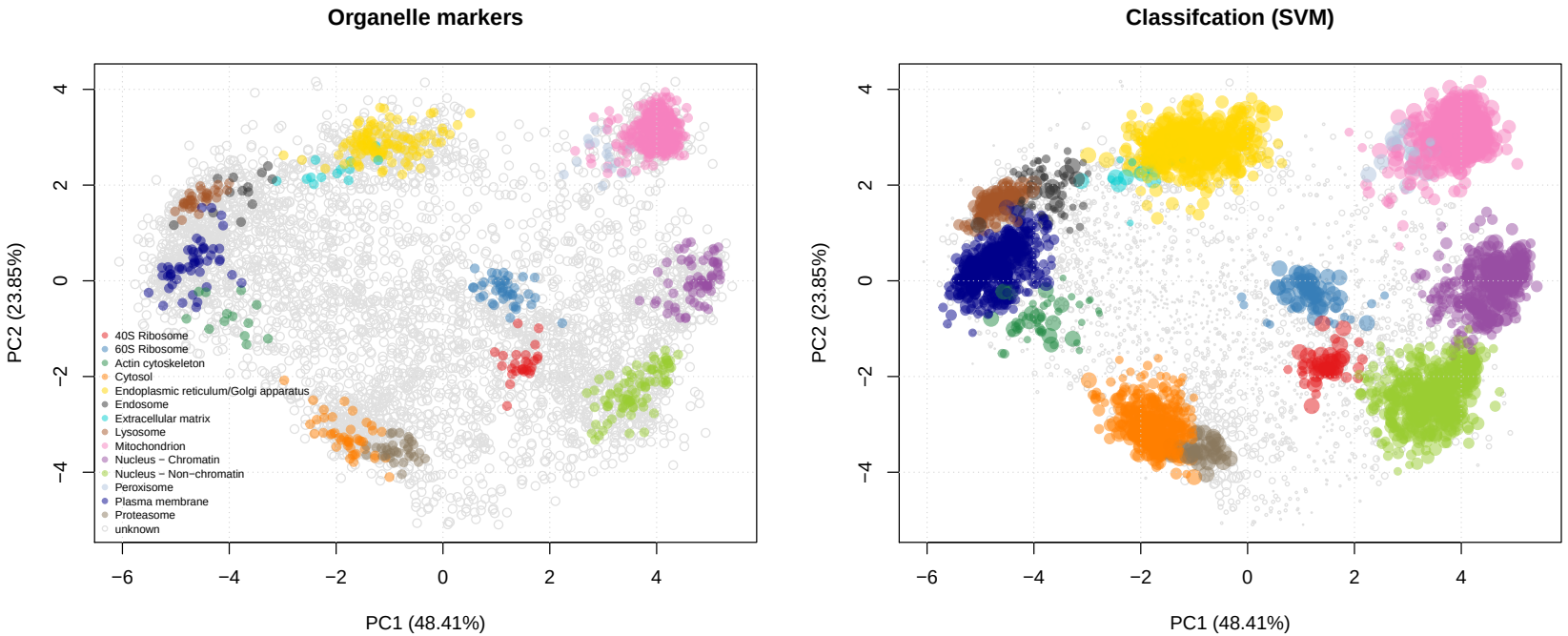
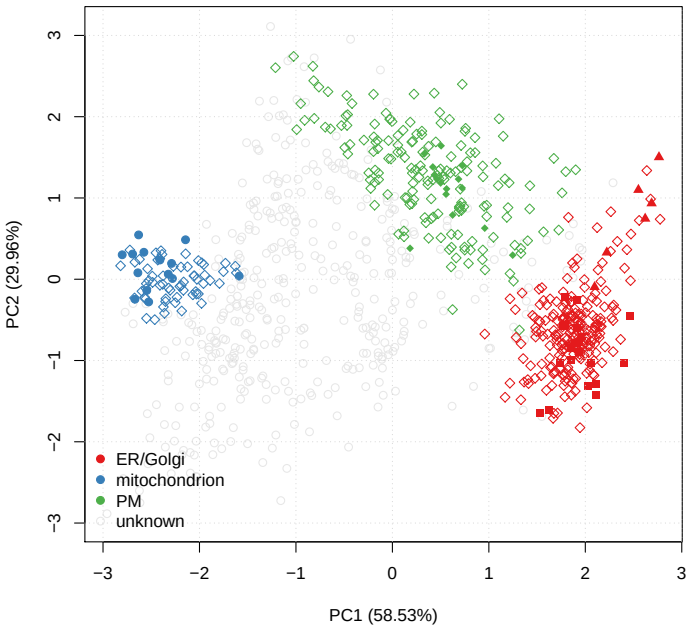
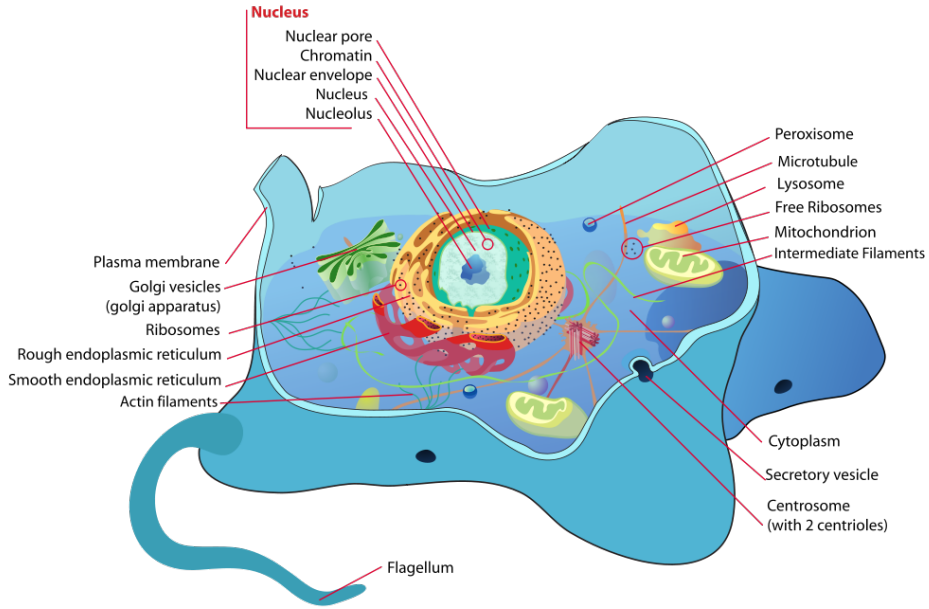


Figure : Support vector machines classifier (after *manually* setting a 5% FDR classification cutoff) on the mouse embryonic stem cell data from Christoforou et al. (2016).

Importance of annotation





Incomplete annotation, and therefore lack of training data, for many/most organelles. *Drosophila* data from Tan et al. (2009).

Semi-supervised learning: novelty detection

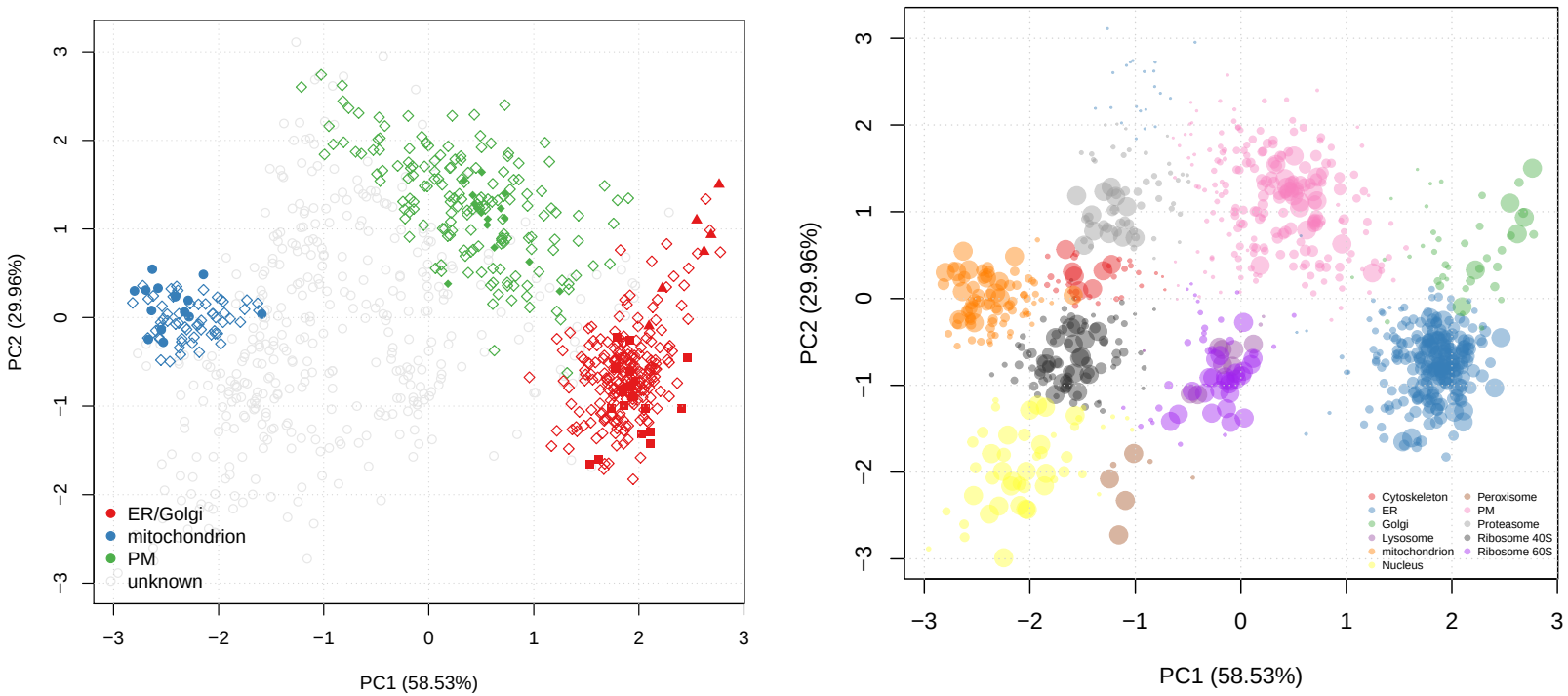


Figure : Left: Original *Drosophila* data from Tan et al. (2009). Right: After semi-supervised learning and classification, Breckels et al. (2013).

Biological discoveries

- Multi-localisation
- Trans-localisation

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*.

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

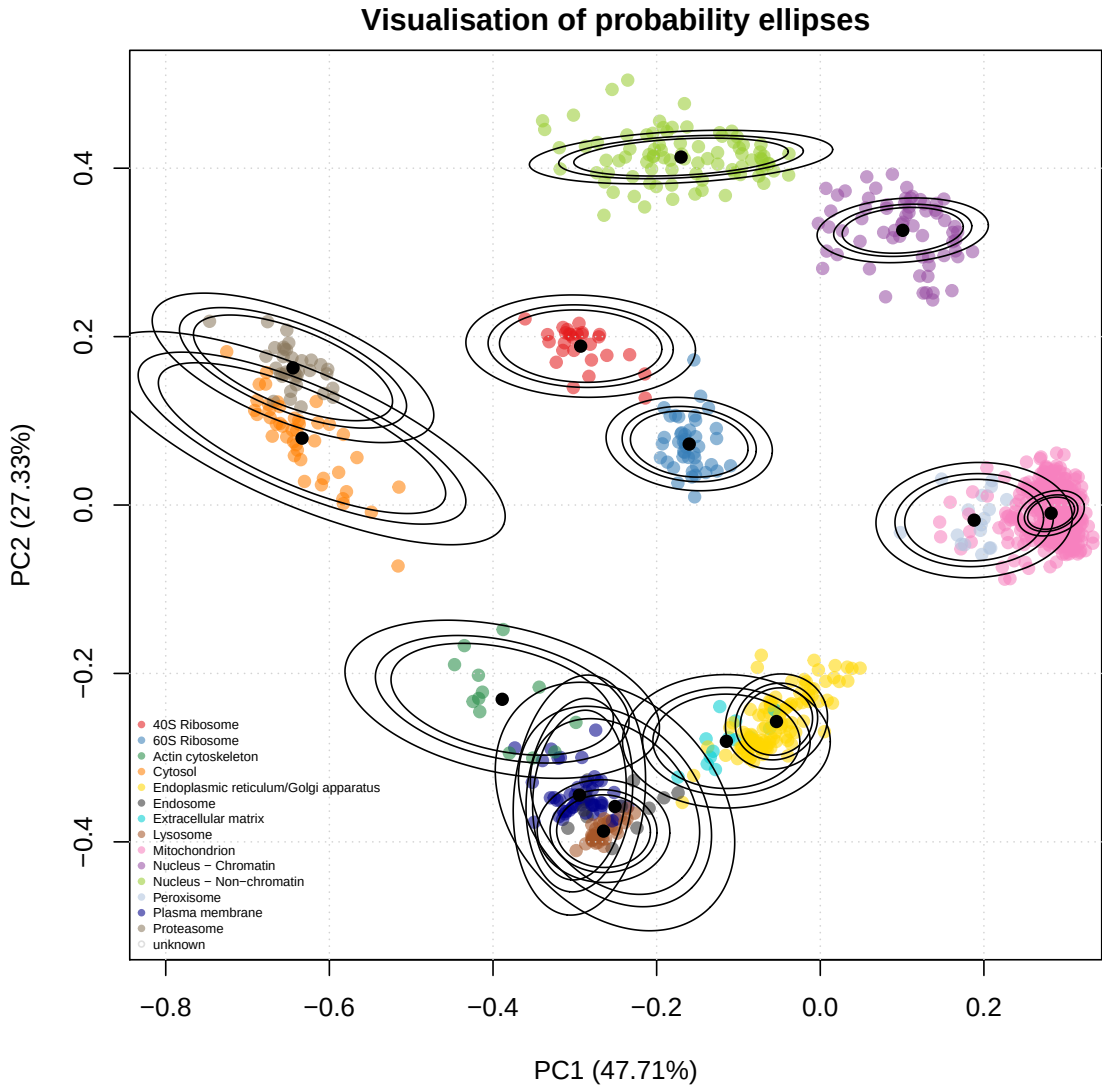


Figure : Illustration of how the TAGM model describes the pluripotent mouse embryonic stem cell data. Each ellipse contains a proportion of total probability of a particular multivariate Gaussian density. The outer ellipse contains 99% of the total probability whilst the middle and inner ellipses contain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

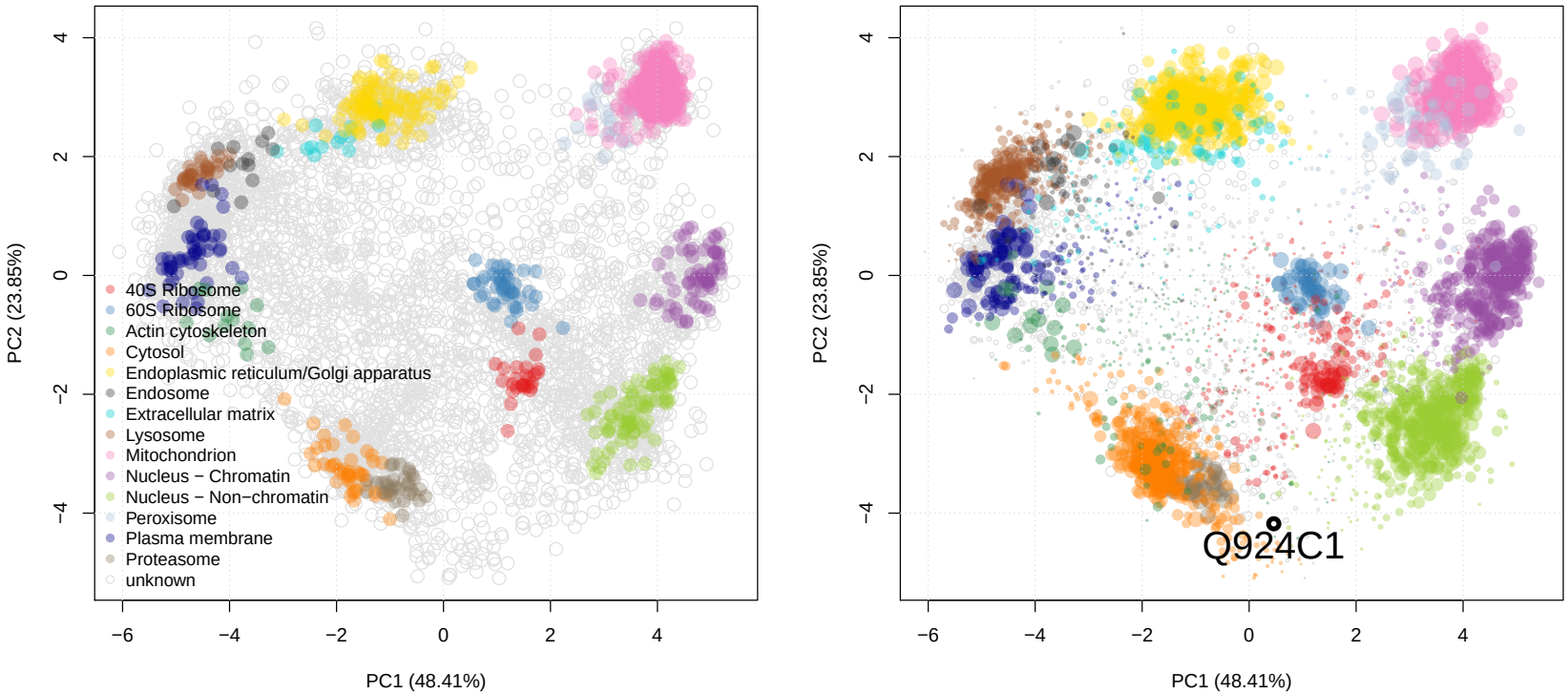
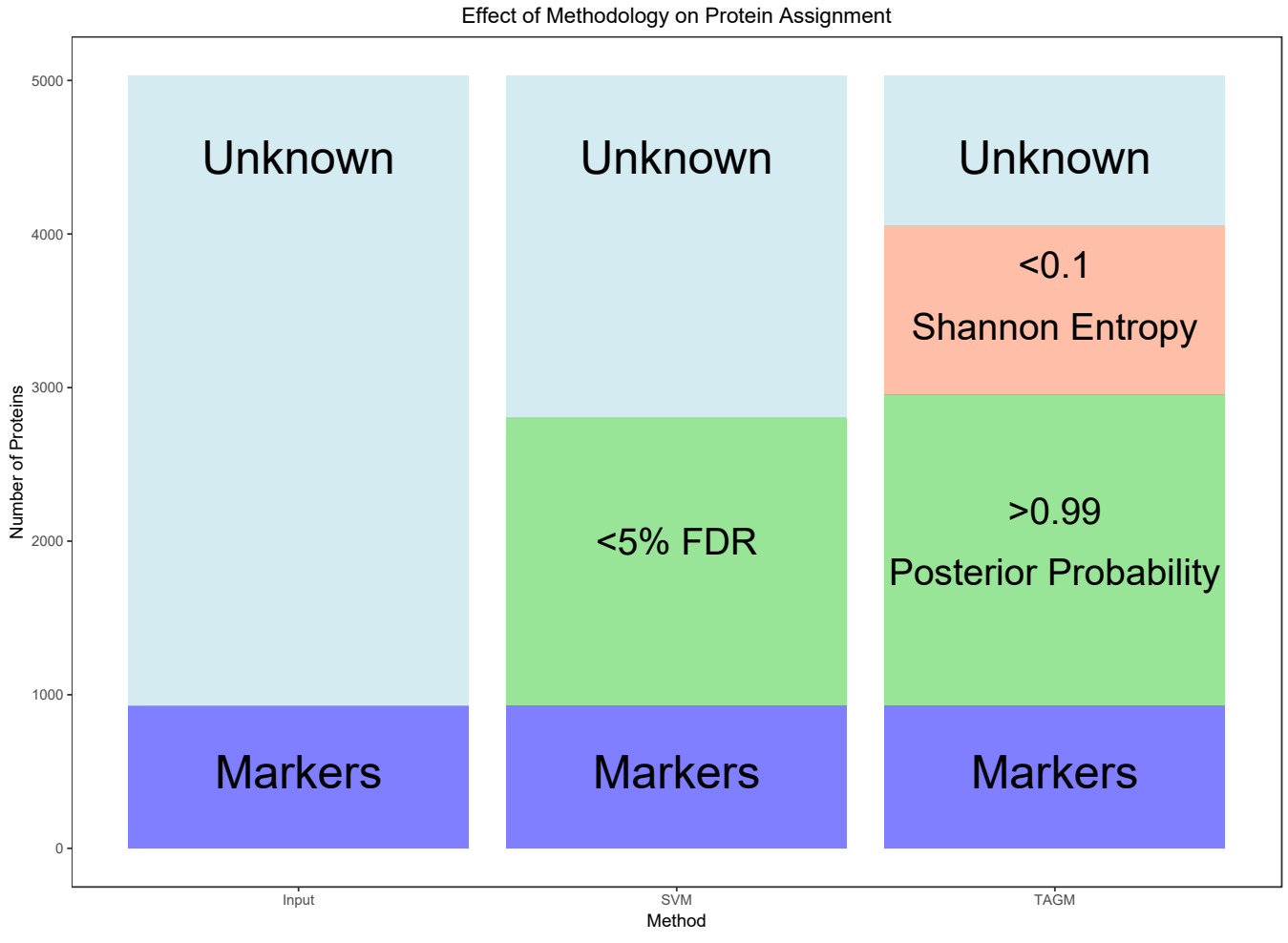


Figure : Assignment of proteins of *unknown* location to one of the annotated classes. The dots are scaled according to the protein assignment probabilities.


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

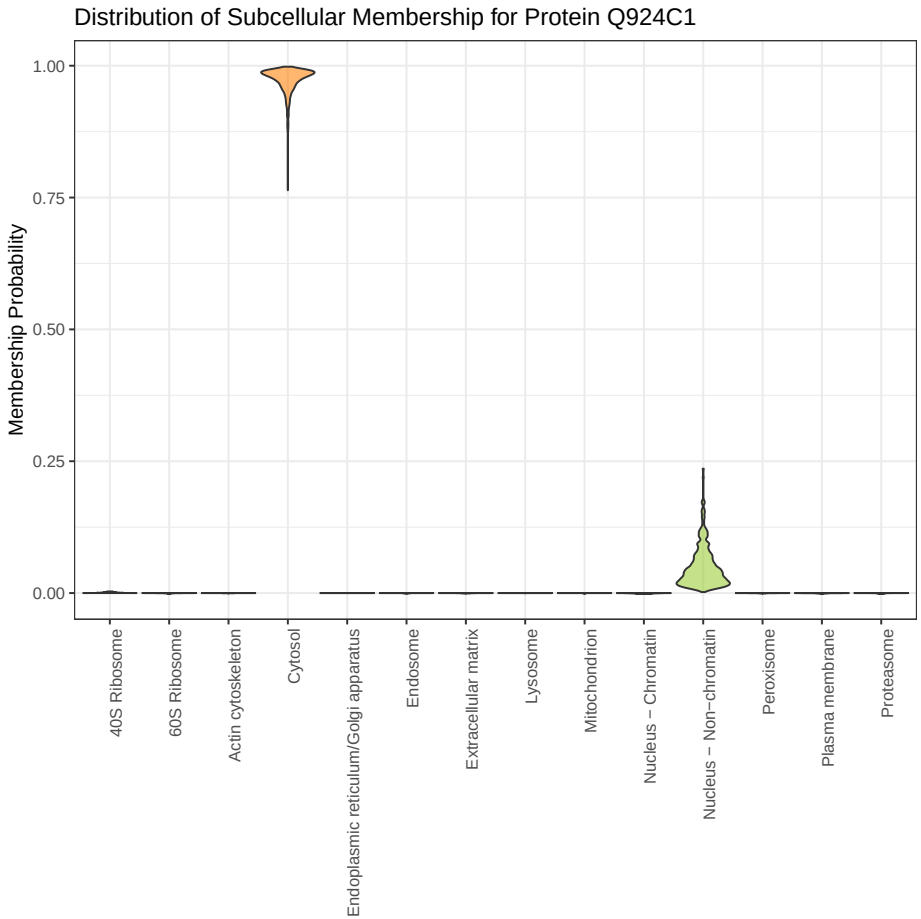
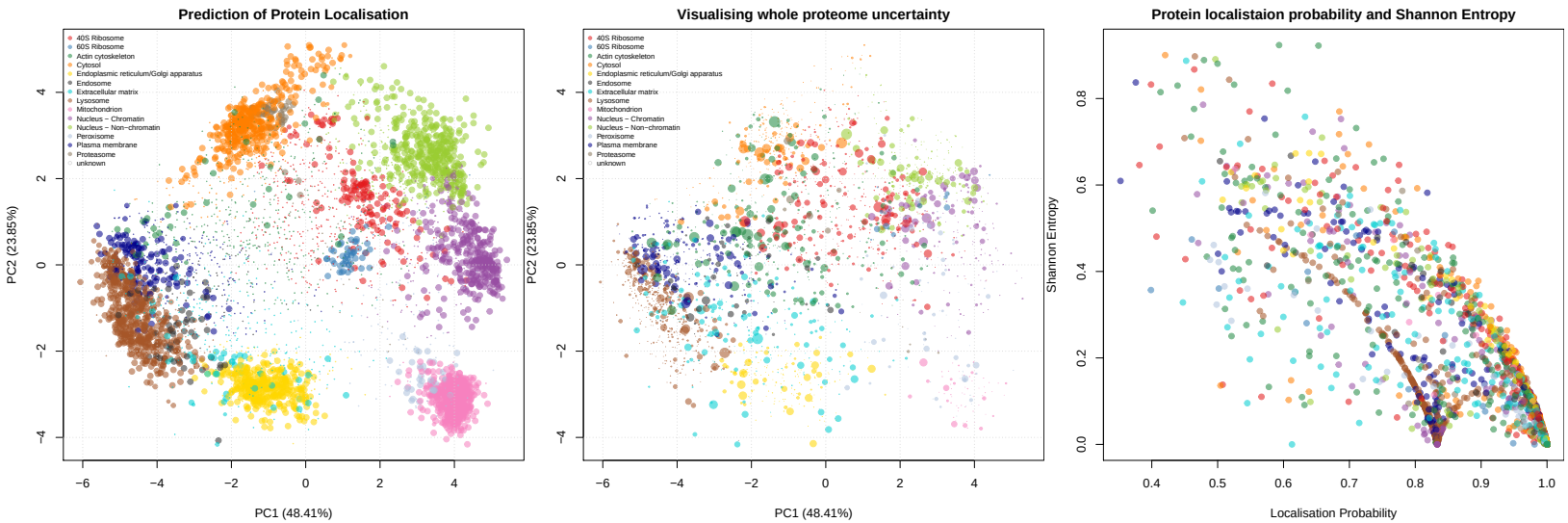


Figure : Exportin 5 (Q924C1) forms part of the micro-RNA export machinery, transporting miRNA from the nucleus to the cytoplasm for further processing. It then translocates back through the nuclear pore complex to return to the nucleus to mediate further transport between nucleus and cytoplasm. The model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

With **Dr Claire Mulvey** at the Cambridge Centre for Proteomics, now at CRUK Cambridge Institute.

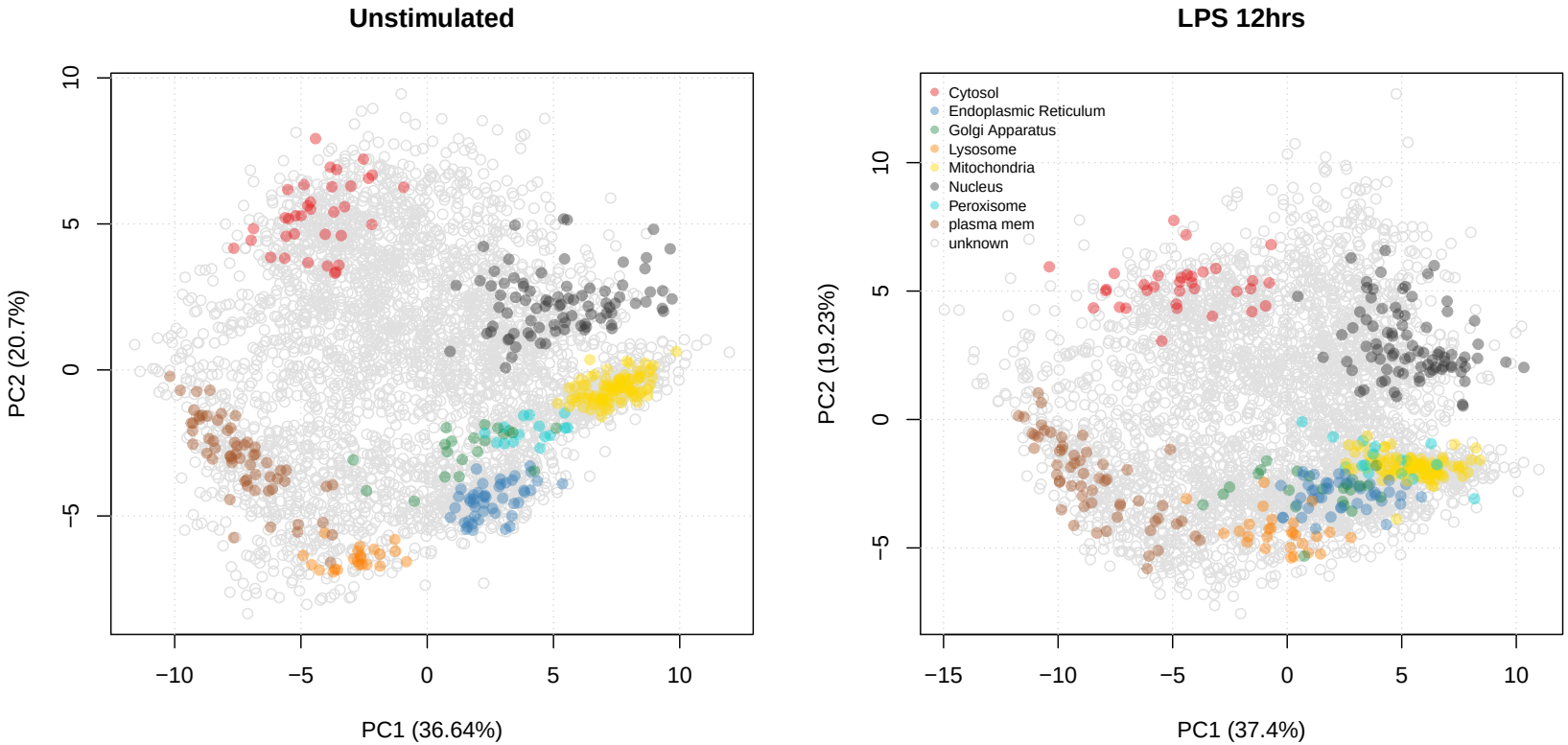


Figure : Spatial maps of unstimulated and LPS-treated cells (combined triplicates).

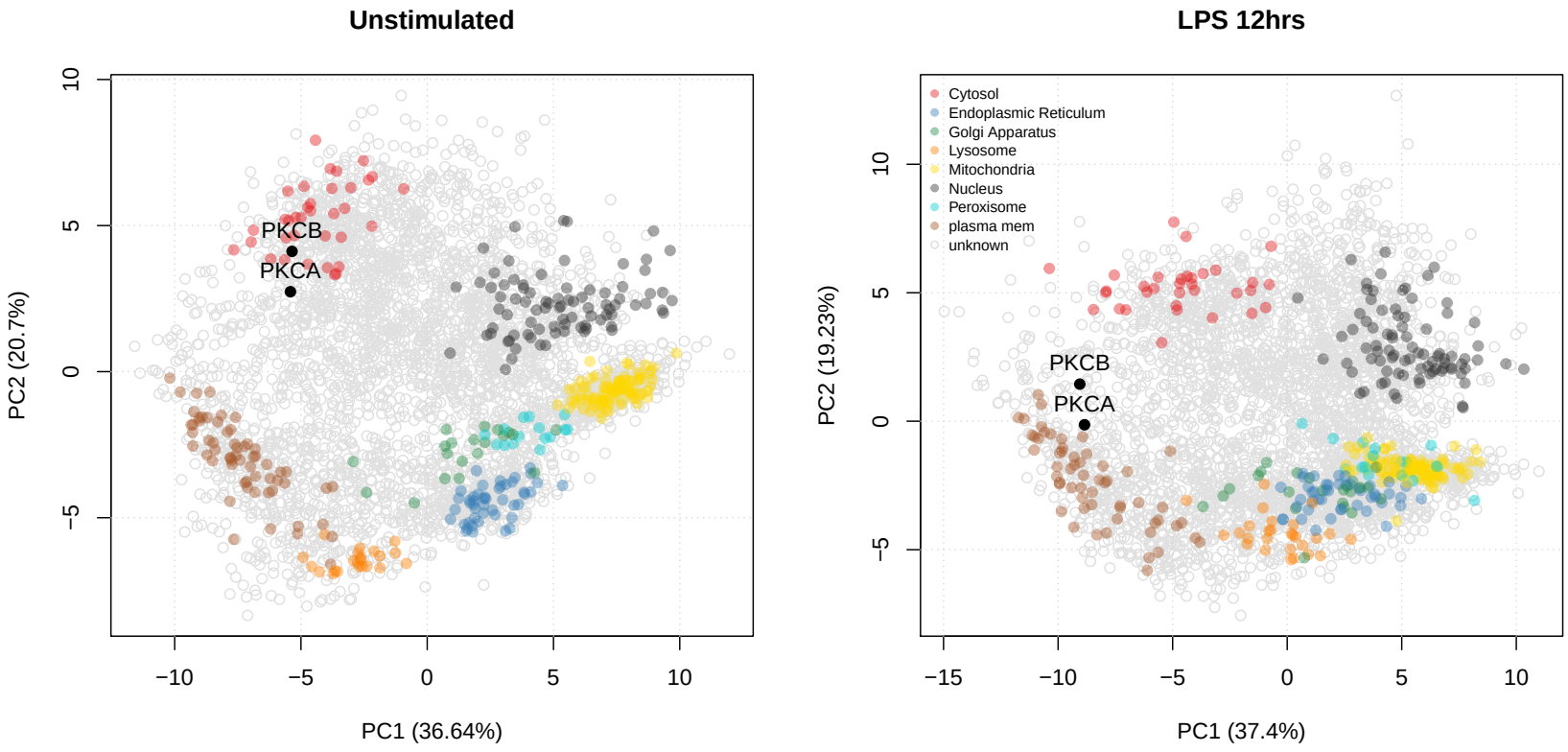


Figure : Relocation of Protein Kinase C α and β from the cytosol to the plasma membrane, **driving maturation into a differentiated macrophage phenotype**.

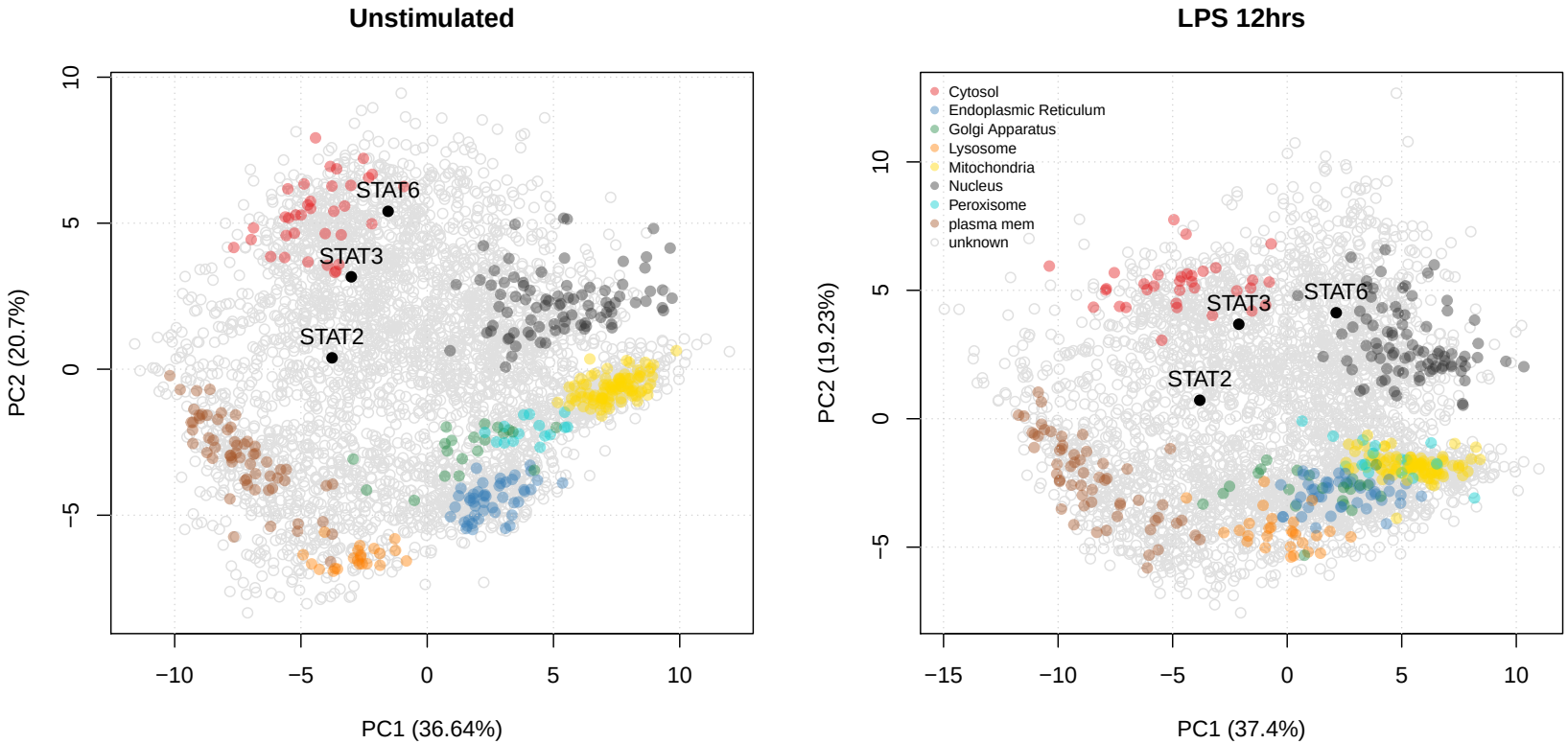


Figure : Relocation of Signal transducer and activator of transcription 6 (STAT6) from the cytosol to the Nucleus, **activating anti-bacterial and anti-viral-like response**. Validated by microscopy and see also Chen et al. (2011).

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/

References I

J Betschinger, J Nichols, S Dietmann, P D Corrin, P J Paddison, and A Smith. Exit from pluripotency is gated by intracellular redistribution of the bhlh transcription factor tfe3. *Cell*, 153(2):335–47, Apr 2013. doi: 10.1016/j.cell.2013.03.012.

L M Breckels, S B Holden, D Wojnar, C M Mulvey, A Christoforou, A Groen, M W Trotter, O Kohlbacher, K S Lilley, and L Gatto. Learning from heterogeneous data sources: An application in spatial proteomics. *PLoS Comput Biol*, 12(5):e1004920, May 2016a. doi: 10.1371/journal.pcbi.1004920.

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.

LM Breckels, L Gatto, A Christoforou, AJ Groen, KS Lilley, and MW Trotter. The effect of organelle discovery upon sub-cellular protein localisation. *J Proteomics*, 88:129–40, Aug 2013.

LM Breckels, CM Mulvey, KS Lilley, and L Gatto. A bioconductor workflow for processing and analysing spatial proteomics data [version 1; referees: awaiting peer review]. *F1000Research*, 5(2926), 2016b. doi: 10.12688/f1000research.10411.1.

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. *bioRxiv*, 2018. doi: 10.1101/282269. URL https://www.biorxiv.org/content/early/2018/03/14/282269.

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.

References II

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.

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.

TR Kau, JC Way, and PA Silver. Nuclear transport and cancer: from mechanism to intervention. *Nat Rev Cancer*, 4(2):106–17, Feb 2004.

K Laurila and M Vihinen. Prediction of disease-related mutations affecting protein localization. *BMC Genomics*, 10:122, 2009.

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.

J E Siljee, Y Wang, A A Bernard, B A Ersoy, S Zhang, A Marley, M Von Zastrow, J F Reiter, and C Vaisse. Subcellular localization of mc4r with adcy3 at neuronal primary cilia underlies a common pathway for genetic predisposition to obesity. *Nat Genet*, Jan 2018. doi: 10.1038/s41588-017-0020-9.

C Stadler, E Rexhepaj, V R Singan, R F Murphy, R Pepperkok, M Uhlén, J C Simpson, and E Lundberg. Immunofluorescence and fluorescent-protein tagging show high correlation for protein localization in mammalian cells. *Nat Methods*, 10(4):315–23, Apr 2013.

DJL Tan, H Dvinge, A Christoforou, P Bertone, A Arias Martinez, and KS Lilley. Mapping organelle proteins and protein complexes in Drosophila melanogaster. *J Proteome Res*, 8(6):2667–2678, Jun 2009.

Acknowledgements

- **Mr Oliver Crook** and **Dr Lisa Breckels**, Computational Proteomics Unit, Cambridge (machine learning, algorithms, software).
- **Dr Sebastian Gibb** and **Dr Johannes Rainer** (software).
- **Prof Kathryn Lilley** *et al.*, Cambridge Centre of Proteomics and **Dr Claire Mulvey**, Cancer Research UK Cambridge Institute (spatial proteomics)
- **Funding**: BBSRC, Wellcome Trust

Slides: https://zenodo.org/record/1435058

Thank you for your attention