Mapping the sub-cellular proteome

Laurent Gatto
lg390@cam.ac.uk – @lgatt0
http://cpu.sysbiol.cam.ac.uk/

Slides: 

https://zenodo.org/record/1180393

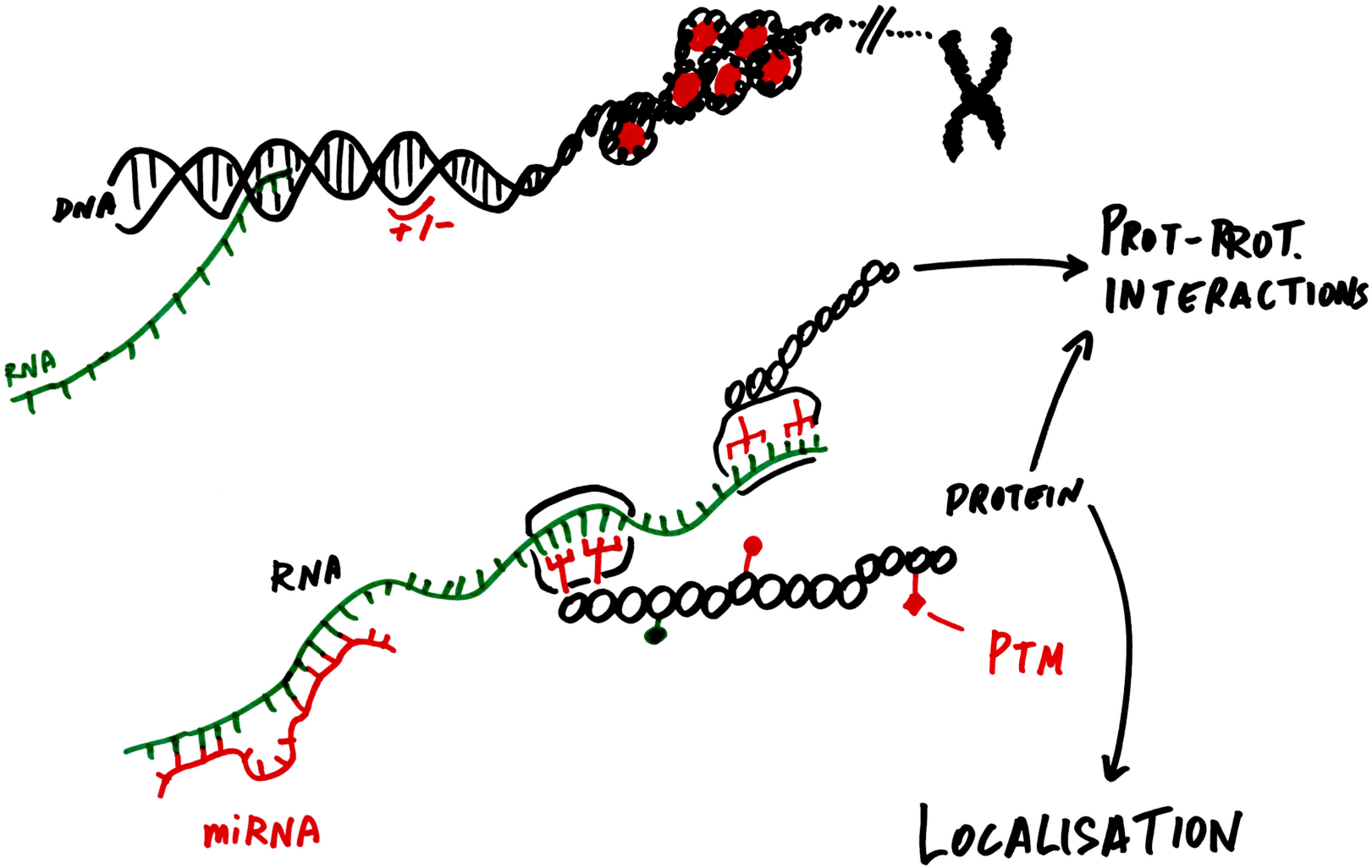
22 Feb 2018, De Duve Institute

Take home messages

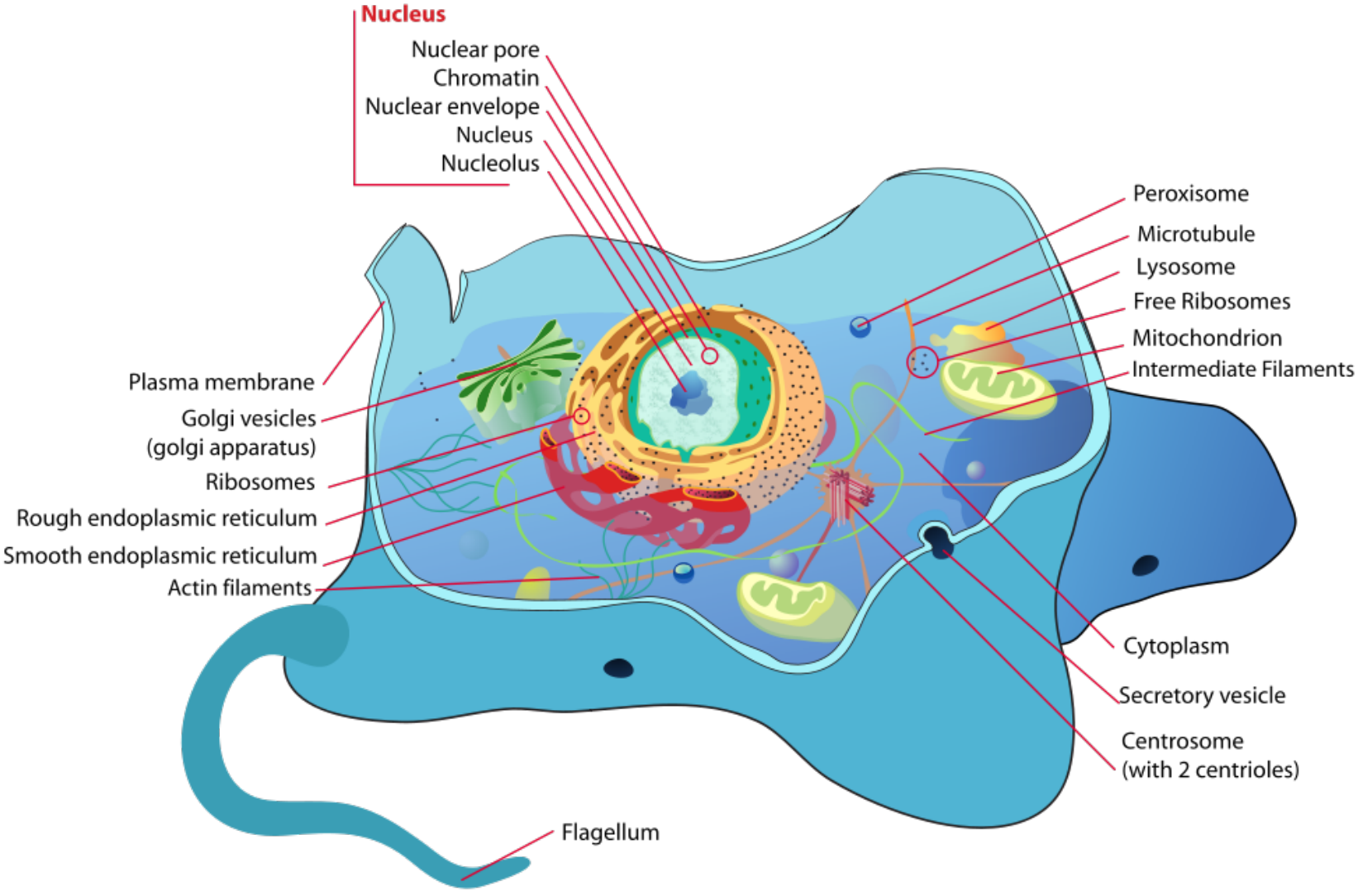
1. Protein sub-cellular localisation: available technologies and opportunities.

2. Reliance on computational biology to acquire reliable biological knowledge.

Regulations



Cell organisation



Spatial proteomics is the systematic study of protein localisations.

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

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.

Examples later.

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

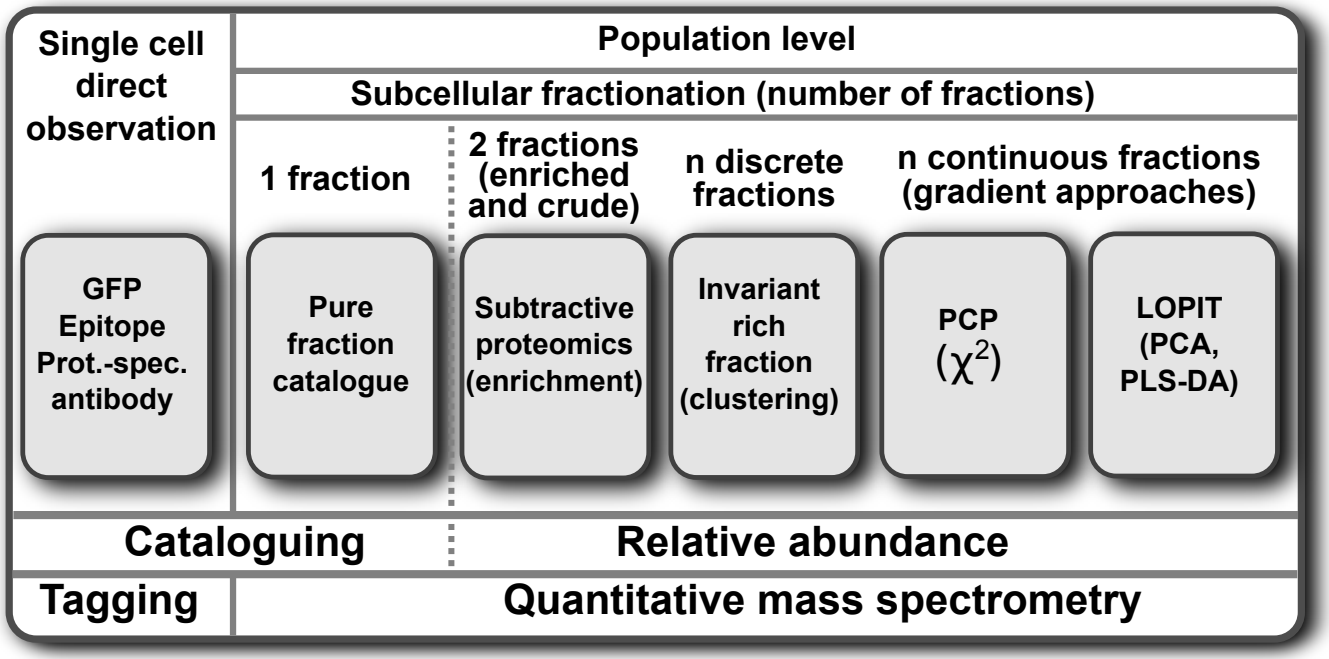


Figure : Organelle proteomics approaches (Gatto et al., 2010)

Fusion proteins and immunofluorescence

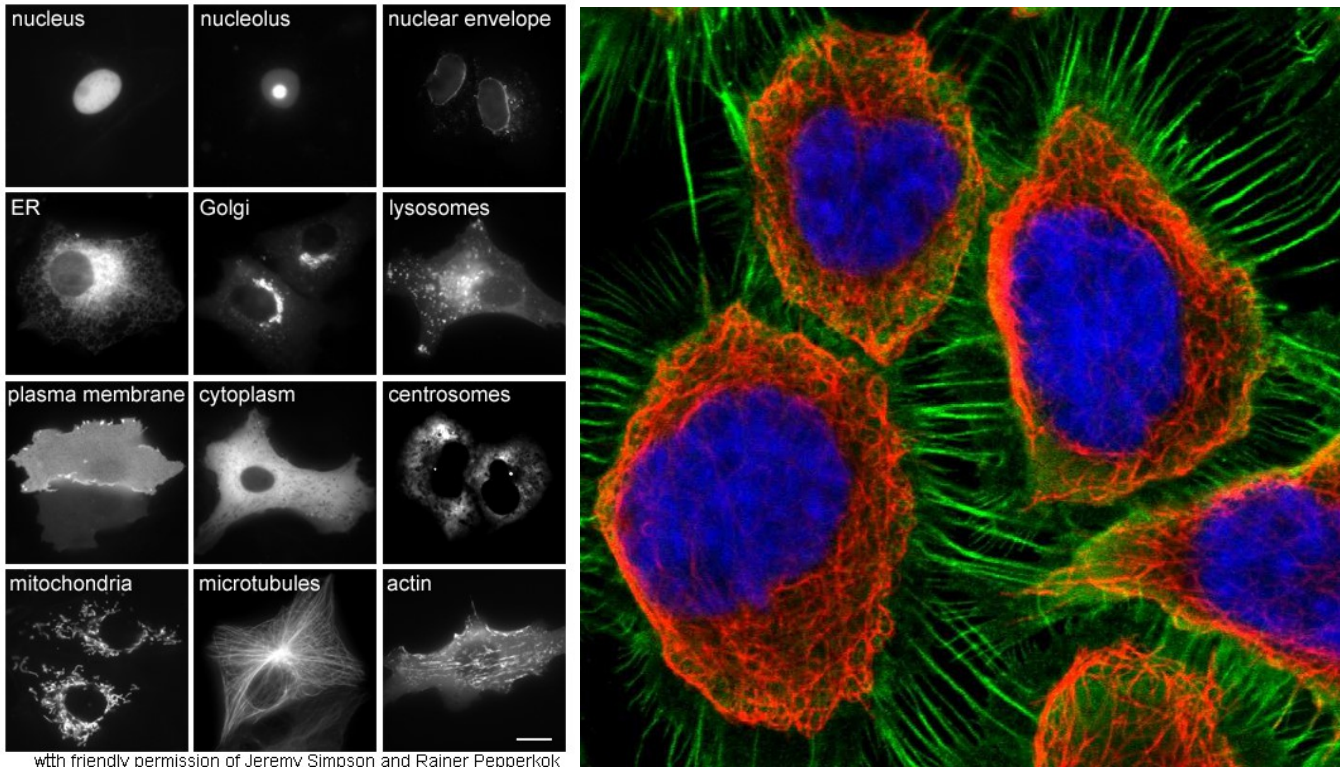


Figure : Targeted protein localisation. Example of discrepancies between IF and FPs as well as between FP tagging at the N and C termini (Stadler et al., 2013).

Spatial proteomics - How, experimentally

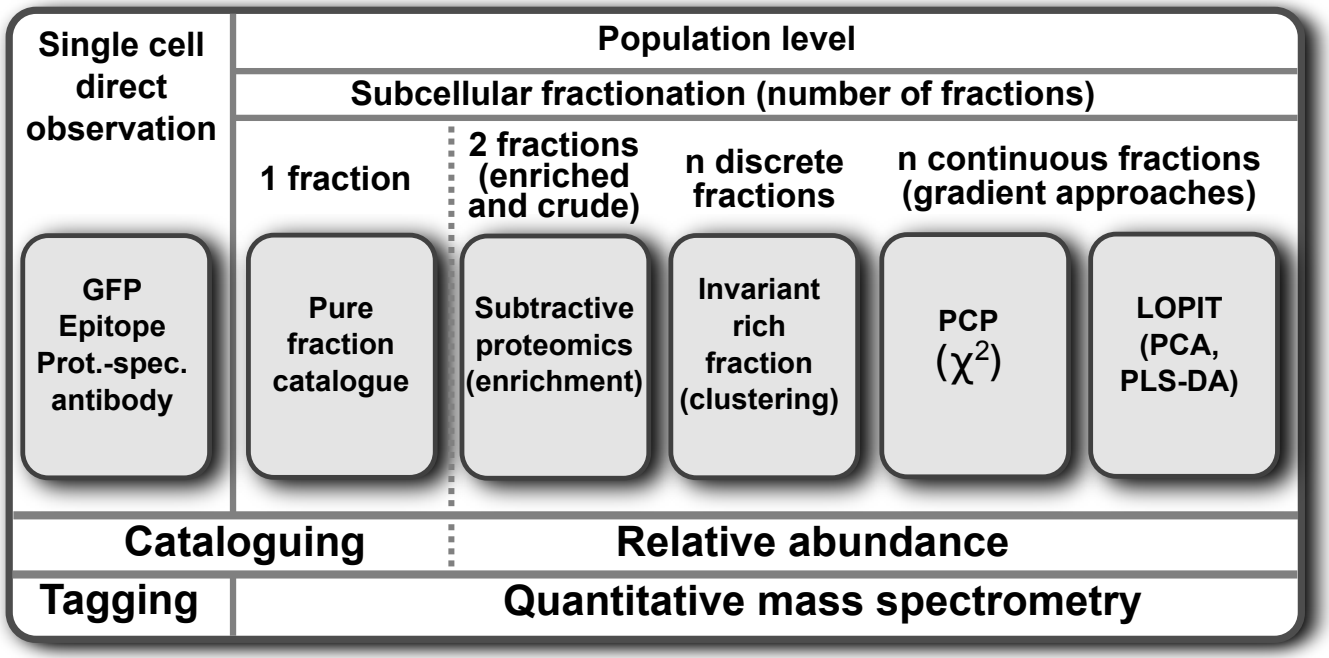
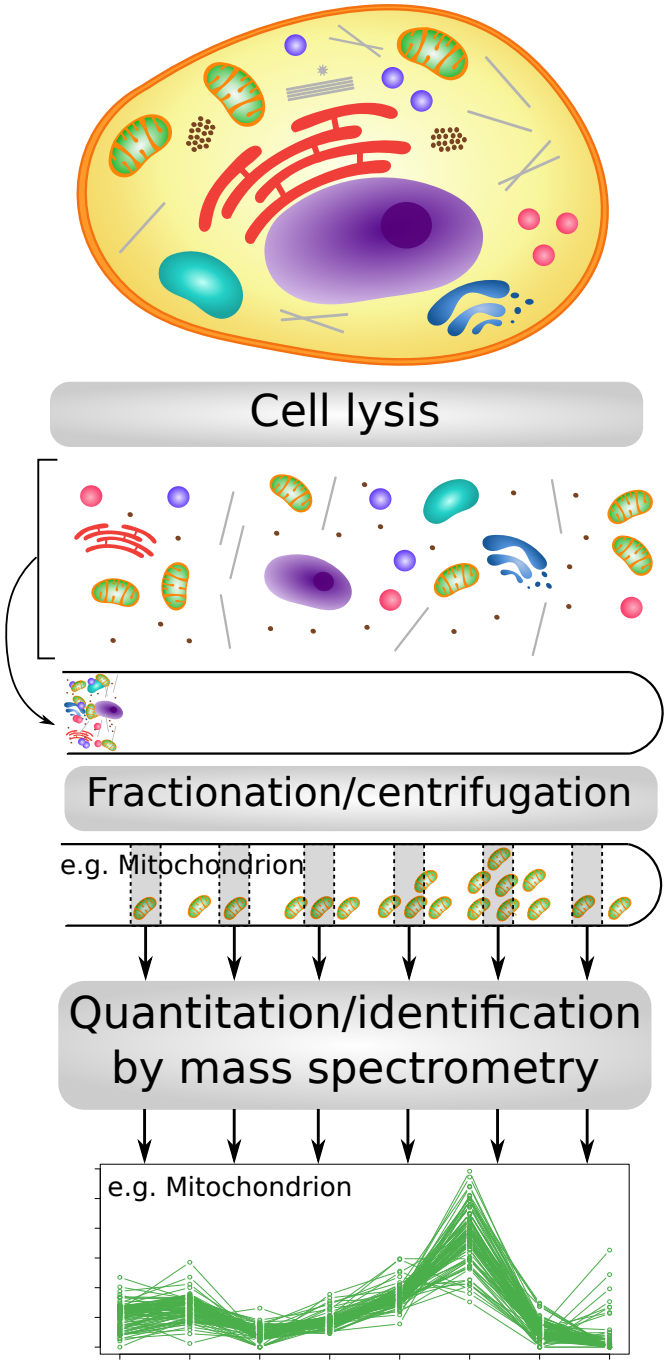


Figure : Organelle proteomics approaches (Gatto et al., 2010).

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 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 (cluster, unsupervised learning)
- Classification (supervised learning)
- Novelty detection (semi-supervised learning)
- Data integration (transfer learning)
- ...

To uncover and understand biology

Visualisation

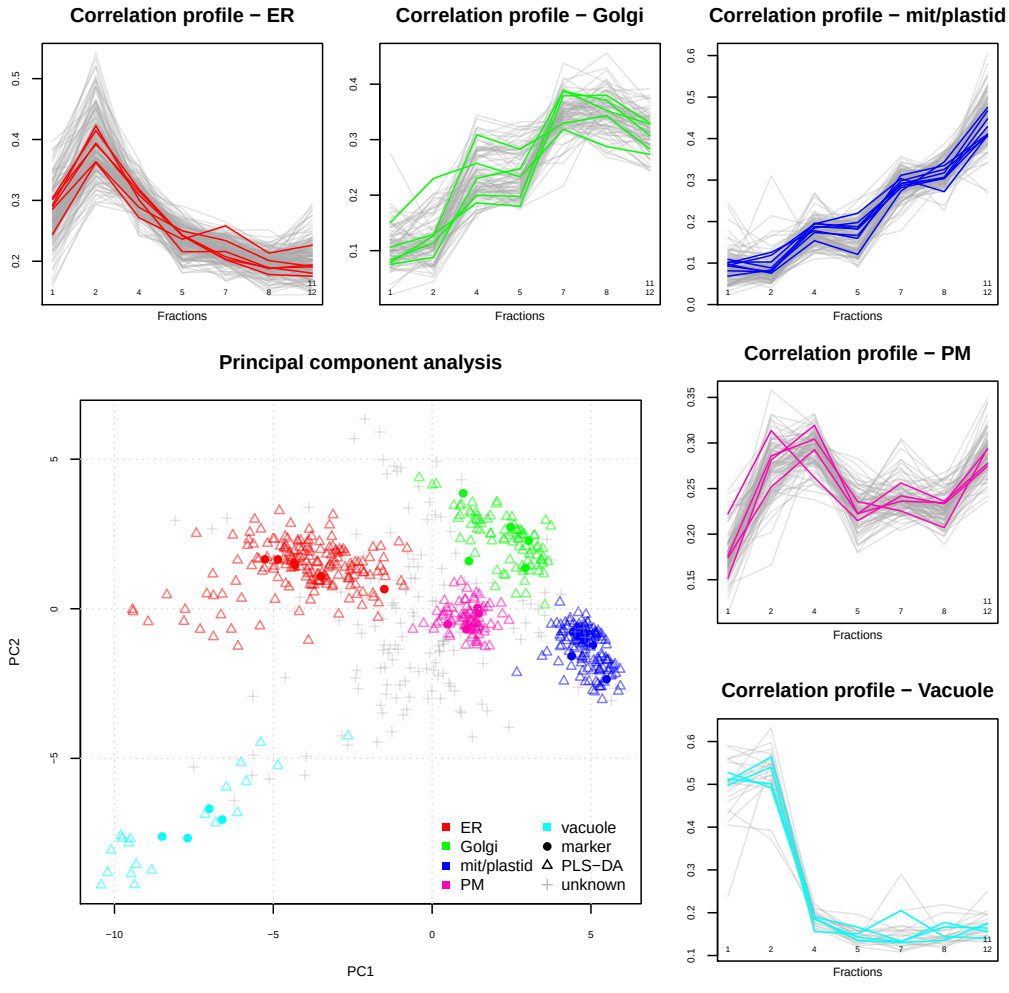


Figure : From Gatto et al. (2010), *Arabidopsis thaliana* data from Dunkley et al. (2006)

Supervised Machine Learning

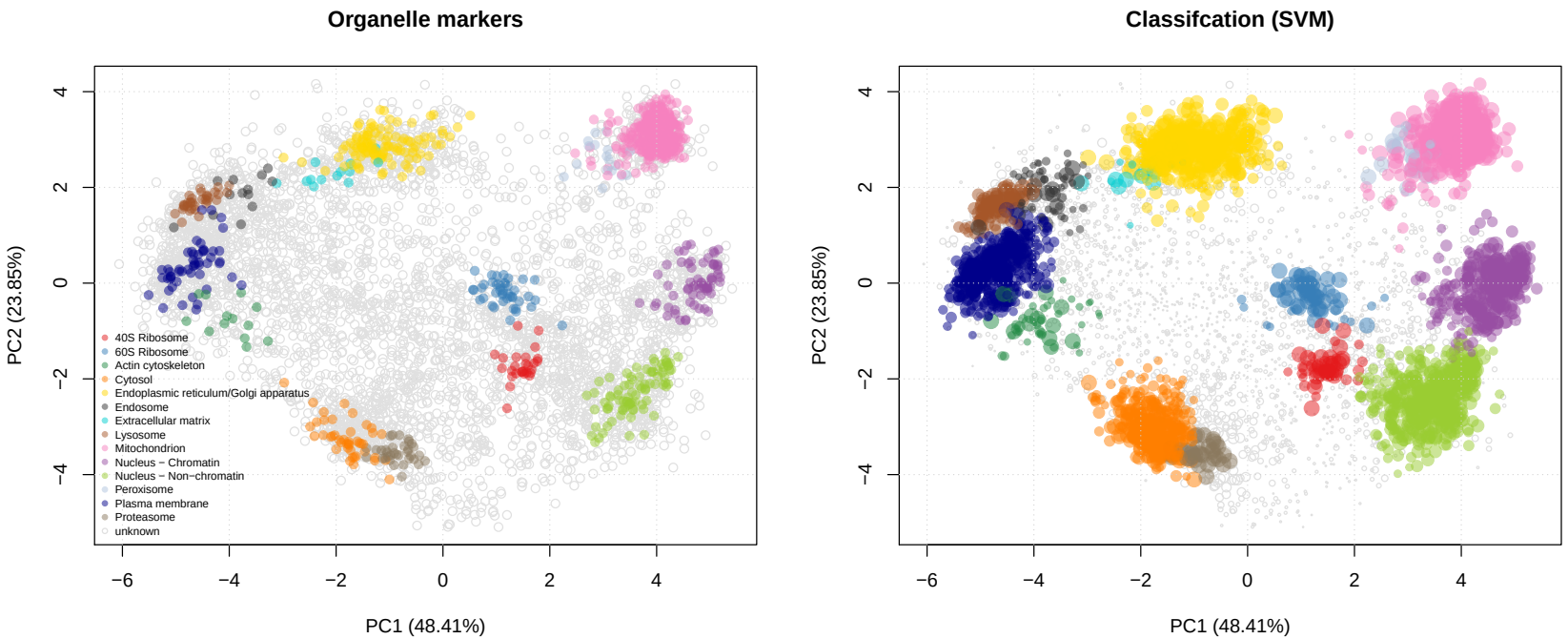
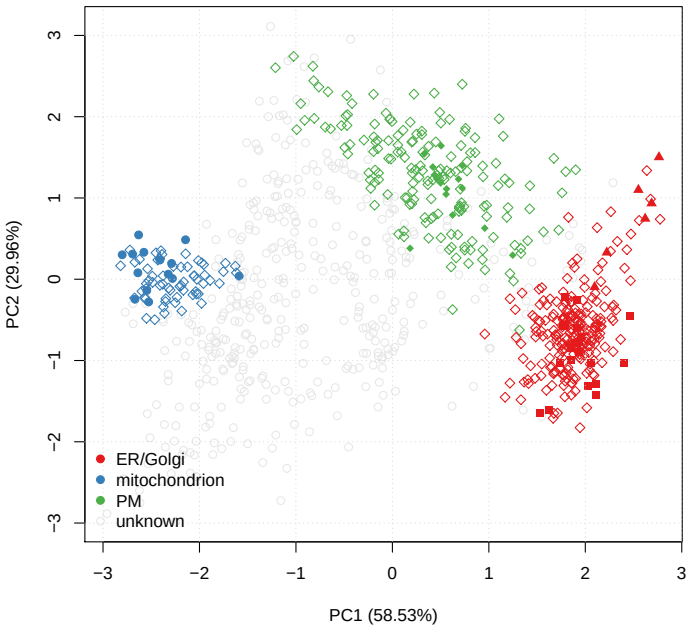
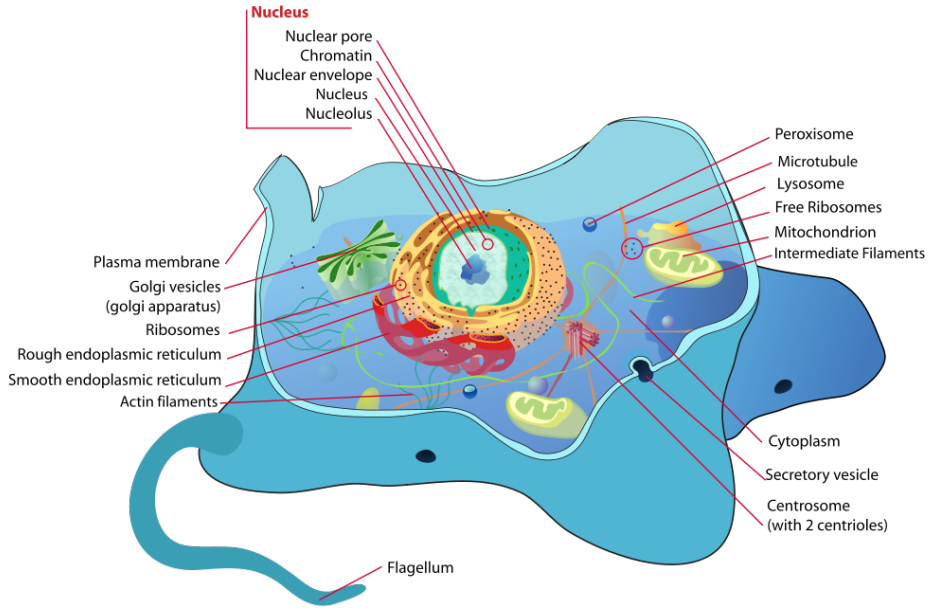


Figure : Support vector machines classifier (after 5% FDR classification cutoff) on the embryonic stem cell data from Christoforou et al. (2016).

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

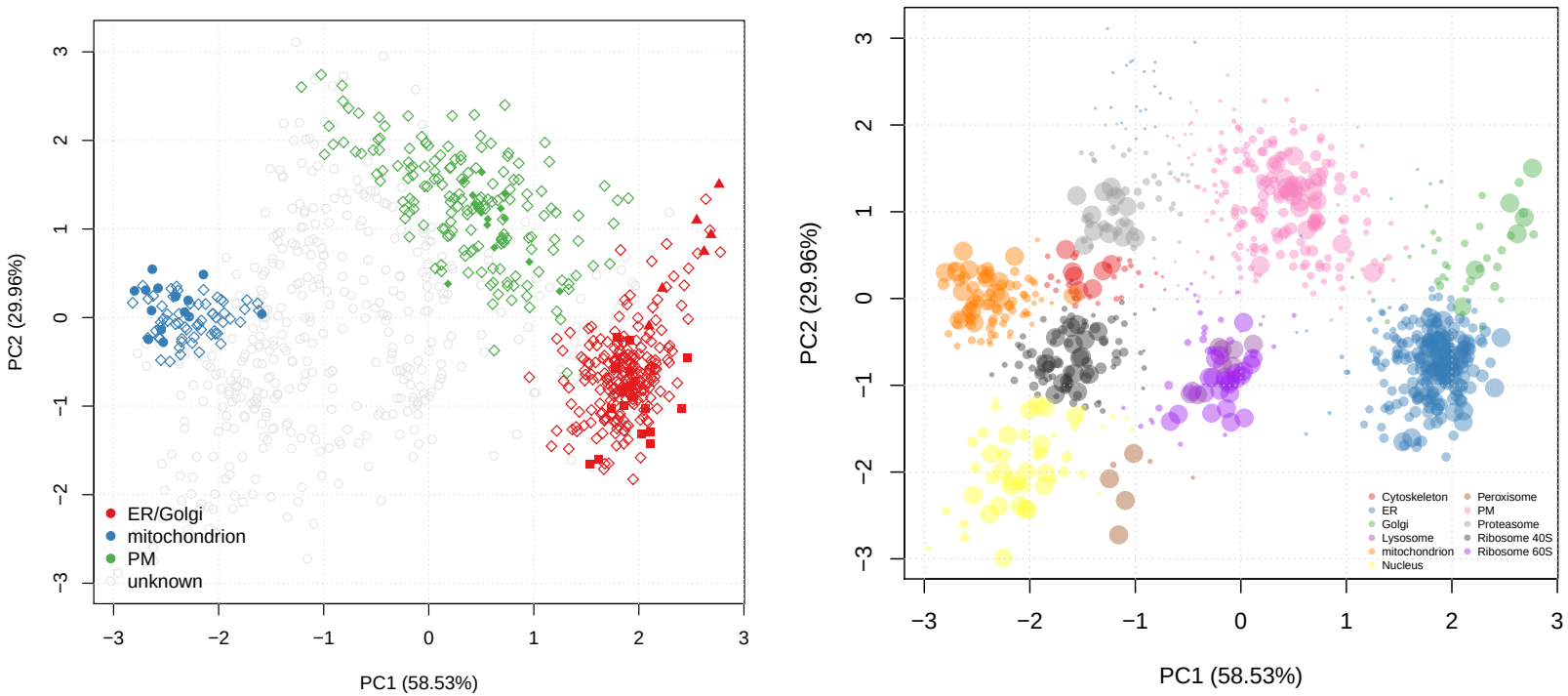


Figure : Left: Original *Drosophila* data from Tan et al. (2009). Right: After semi-supervised learning and classification, Breckels et al. (2013).

Improving on LOPIT

Improving is obtaining better **sub-cellular resolution** to increase the number of protein that can be **confidently** assigned to a sub-cellular niche $\Rightarrow$ **biological discoveries**.

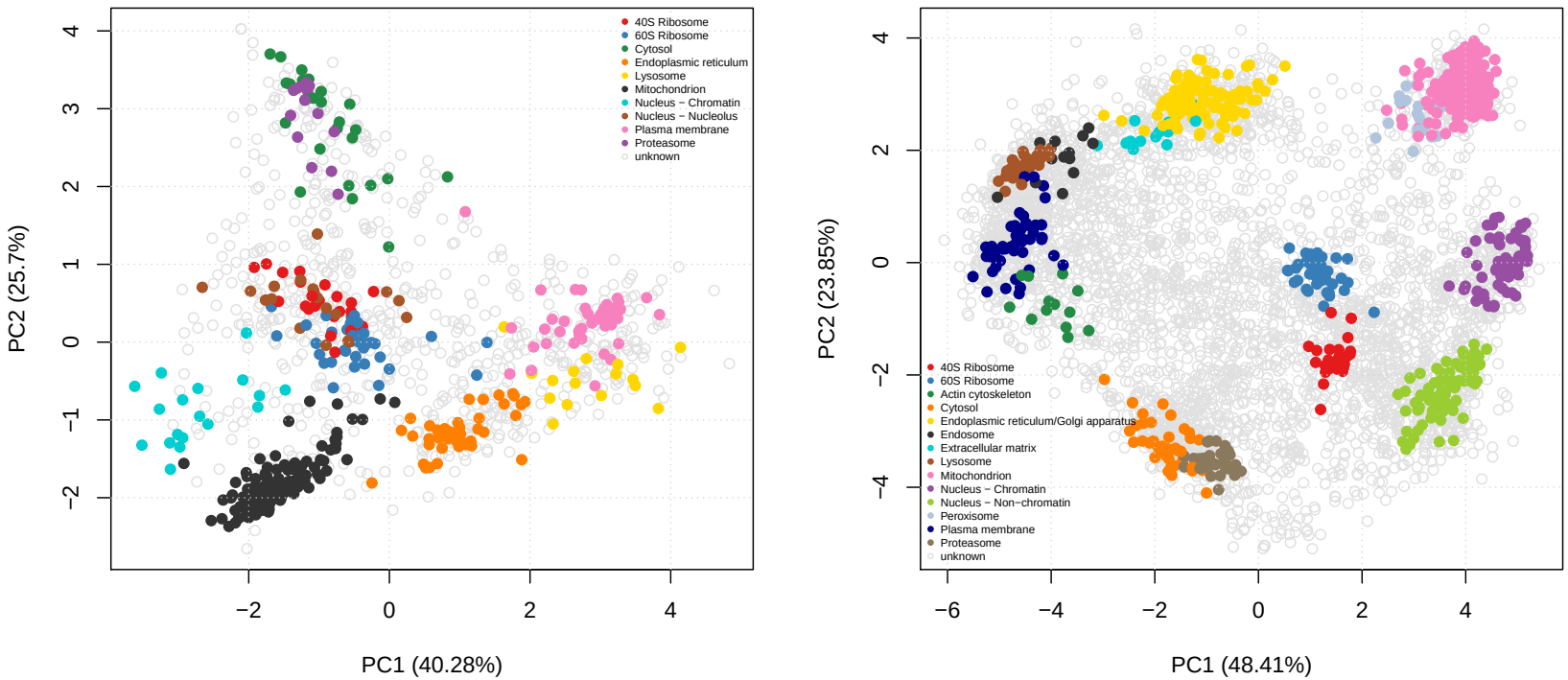


Figure : E14TG2a embryonic stem cells: old (left, published in Breckels et al. (2013)) *vs.* new, better resolved (right) experiments (Christoforou et al. (2016)).

Improving on LOPIT

LOPIT Dunkley et al. (2006) Gatto et al. (2014a)	**Computational**: *transfer learning* Breckels et al. (2016a)
Experimental: *hyperLOPIT* Christoforou et al. (2016) Mulvey et al. (2017) Breckels et al. (2016b)	**Biological discoveries**

Experimental advances: hyperLOPIT Christoforou et al. (2016)

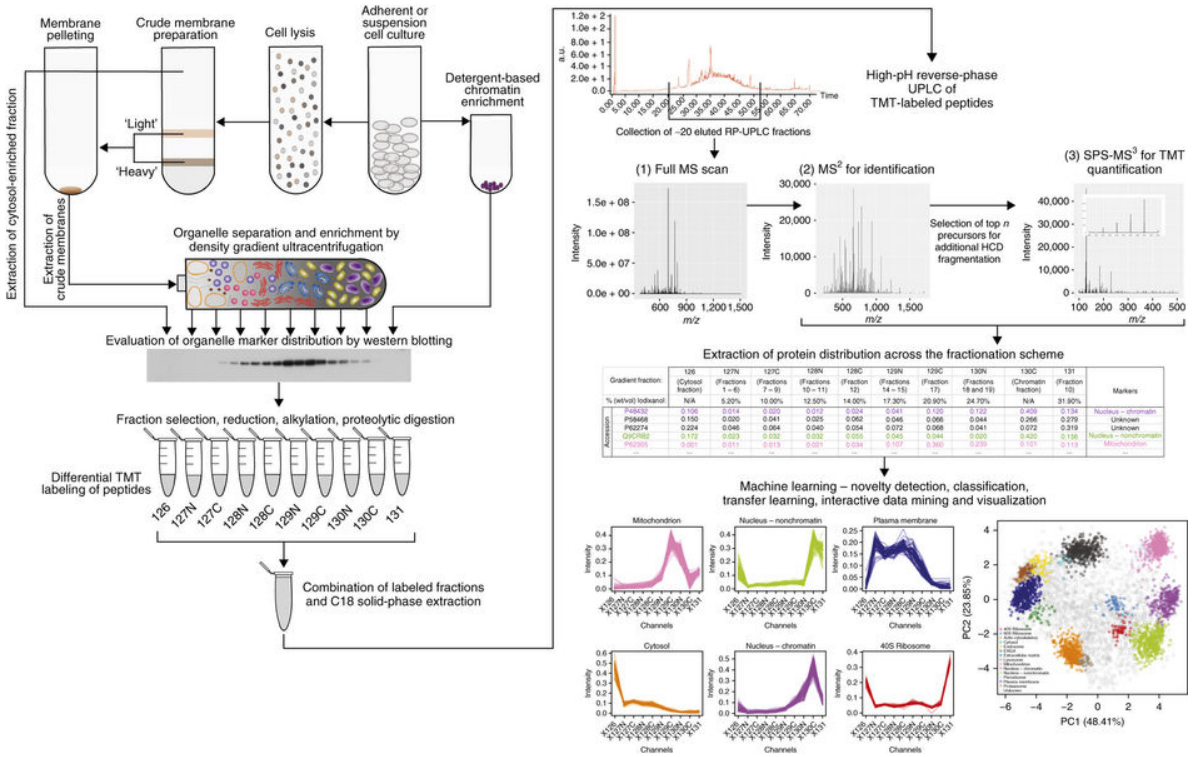


Figure : From Mulvey et al. (2017) *Using* **hyperLOPIT** *to perform high-resolution mapping of the spatial proteome*: (1) organelle separation and enrichment by **density gradient ultracentrifugation**, (2) **chromatin and cytosol** enrichment fractions, and (3) accurate quantification using **synchronous precursor selection (SPS)-MS^3 for TMT 11-plex** quantification.

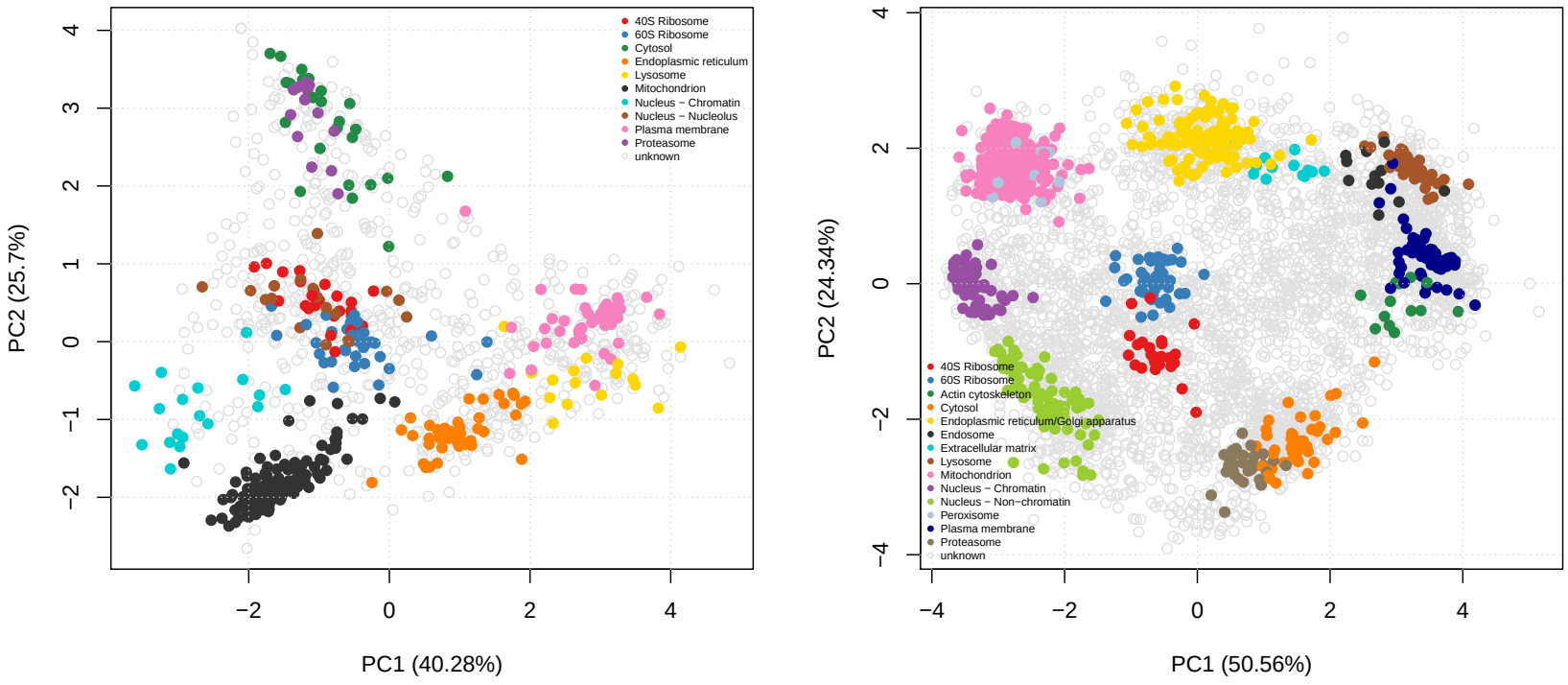


Figure : E14TG2a LOPIT on 8 fractions (using iTRAQ 8-plex) and 1109 proteins *vs.* hyperLOPIT on 10 fractions (using TMT 11-plex) and SPS-MS^3 for 5032 proteins.

Computational advances: Transfer learning

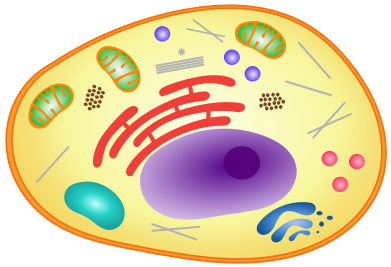
What about using **addition data**, such as annotations from the Gene Ontology (GO), sequence features (pseudo aminoacid composition), signal peptide, trans-membrane domains (length, number, ...), images (IF, FP), interaction data, prediction software, ...

- From a user perspective: **"free/cheap"** vs. expensive and time-consuming experiments.
- Abundant (all proteins, 100s of features) vs. (experimentally) limited/**targeted** (1000s of proteins, 6 – 20 of features)
- For localisation in system at hand: *low* vs. high **quality**
- **Static** vs. **dynamic**

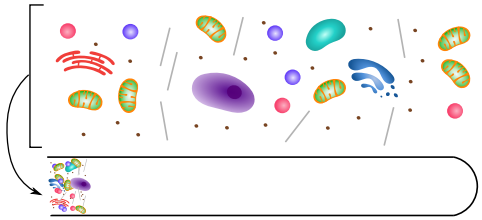
Transfer learning

Support/complement the **primary** target domain (experimental data) with **auxiliary** data (annotation, imaging, PPI, ...) features without compromising the integrity of our primary data.

PRIMARY EXPERIMENTAL DATA

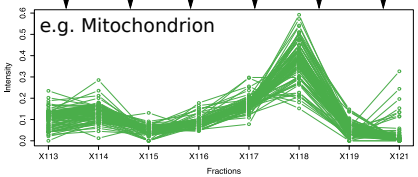


Cell lysis



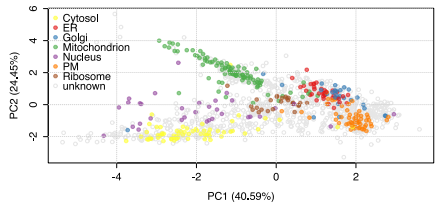
Fractionation/centrifugation

Quantitation/identification by mass spectrometry



	X113	X114	X115	X116	X117	X118	X119	X121
O00767	0.1361	0.150	0.1062	0.147	0.277	0.1429	0.0380	0.00338
P51648	0.1914	0.205	0.0566	0.165	0.237	0.0996	0.0180	0.02727
Q2TAA5	0.1297	0.201	0.0546	0.146	0.292	0.1463	0.0206	0.00902
Q9UKV5	0.0939	0.207	0.0419	0.204	0.344	0.1098	0.0000	0.00000

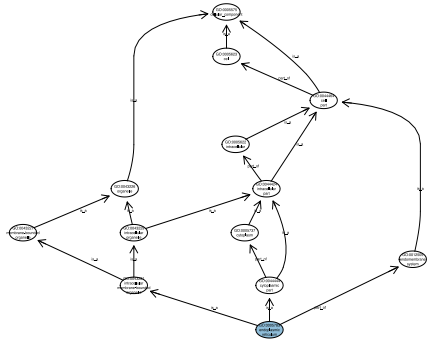
Visualisation



AUXILIARY DRY DATA

Database query

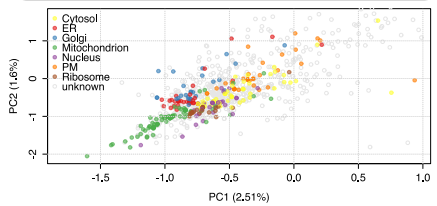
Extract GO CC terms



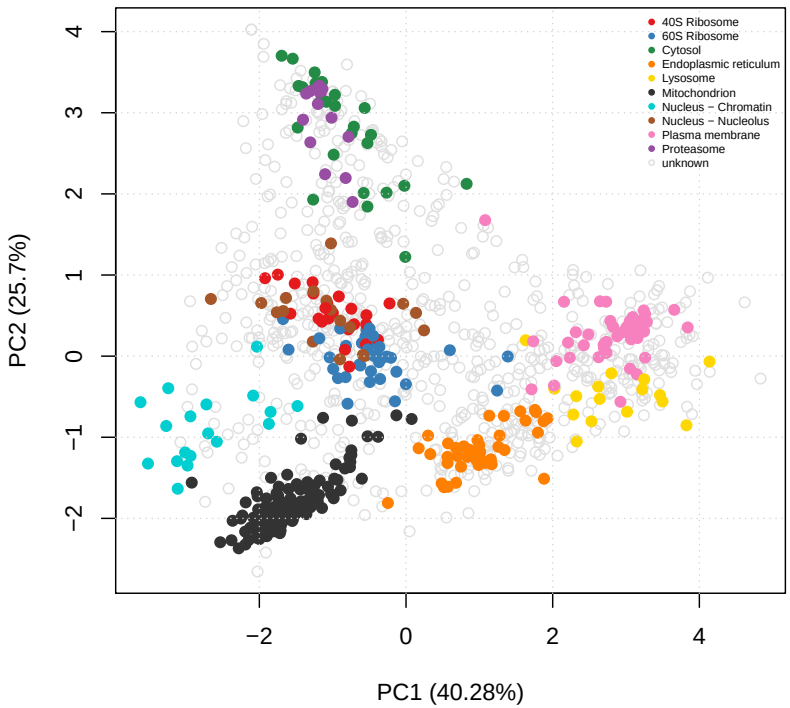
Convert terms to binary

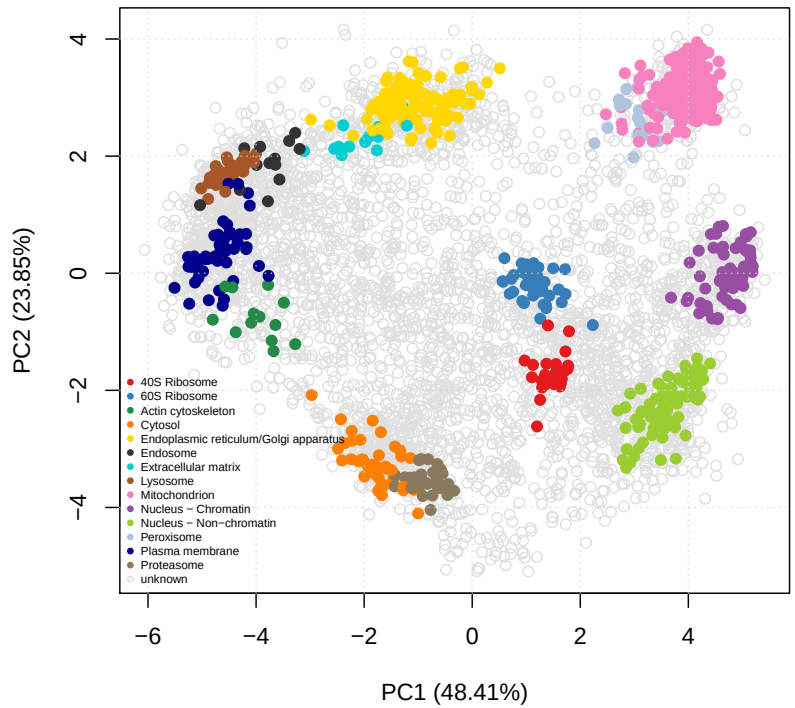
	GO:0016021	GO:0005789	GO:0005783	...	...	...
O00767	1	1	1	...	...	...
P51648	1	1	0	...	...	...
Q2TAA5	1	1	0	...	...	...
Q9UKV5	0	0	0	...	...	...

Visualisation



Breckels et al. (2016a) *Learning from Heterogeneous Data Sources: An Application in Spatial Proteomics.*





Application of **transfer learning** on the *old* **E14TG2a** embryonic stem cells (left, Breckels et al. (2013)) and **GO cellular compartment**, and validated using the *new*, better resolved, hyperLOPIT data (right, Christoforou et al. (2016)).

Transfer learning results

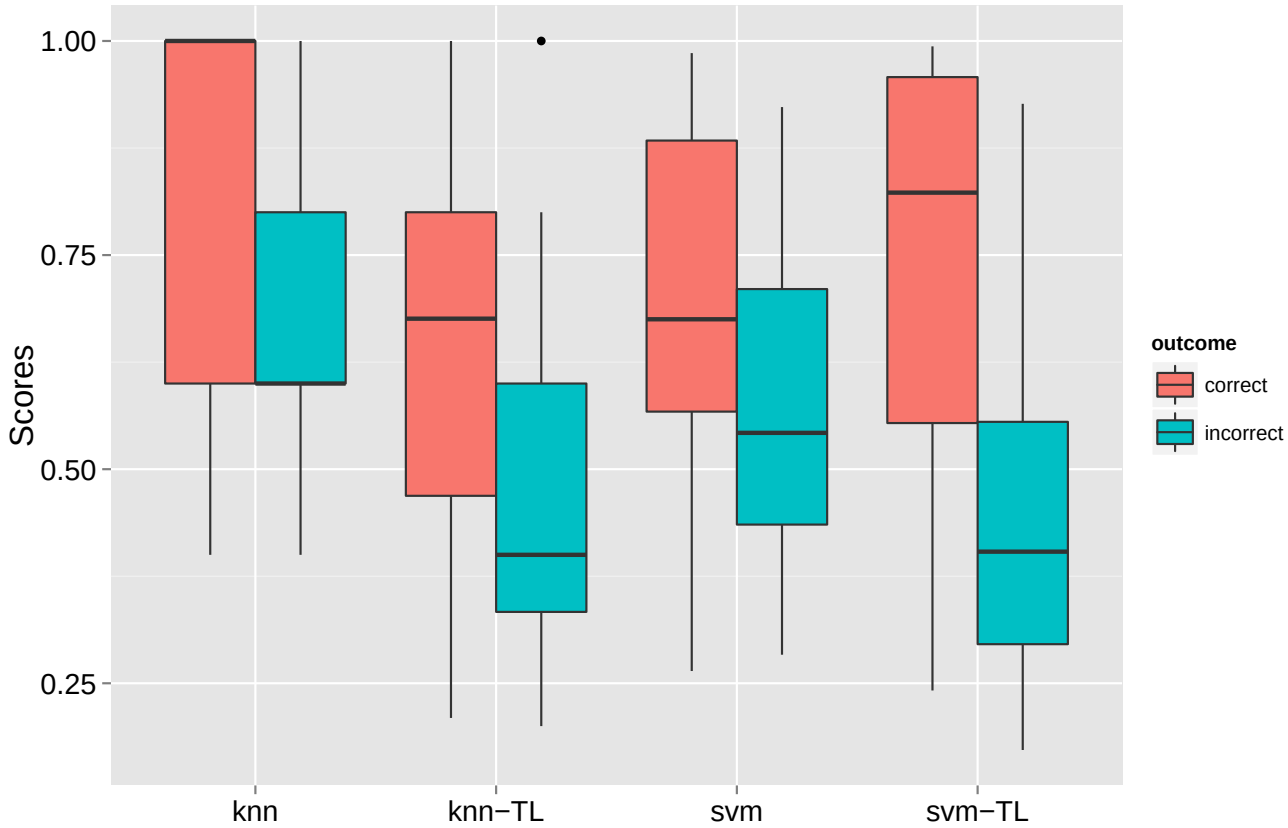


Figure : From Breckels et al. (2016a) *Learning from heterogeneous data sources: an application in spatial proteomics*.

Biological discoveries

- Multi-localisation
- Trans-localisation

Dependent on good sub-cellular resolution and **adequate computational tools.**

Embracing uncertainty

A Bayesian Mixture Modelling Approach For Spatial Proteomics

We propose a Bayesian generative classifier based on Gaussian mixture models to assign proteins probabilistically to sub-cellular niches, thus proteins have a probability distribution over sub-cellular locations.

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

Embracing uncertainty

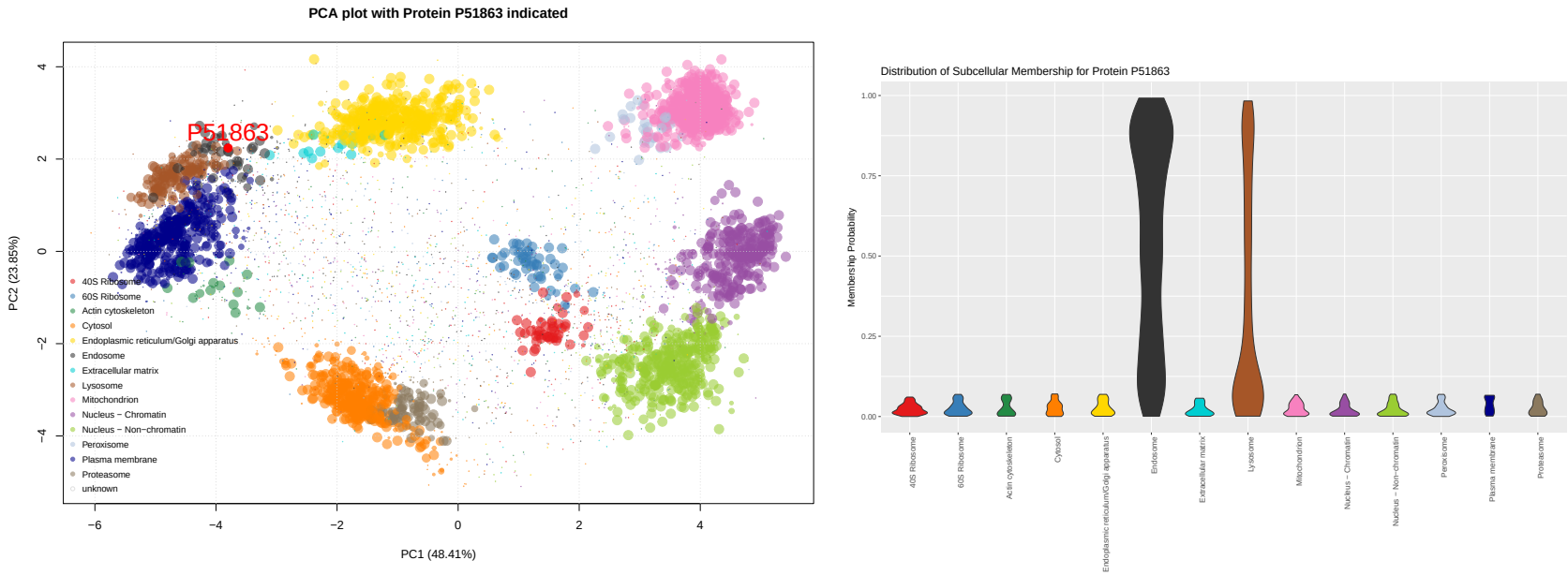
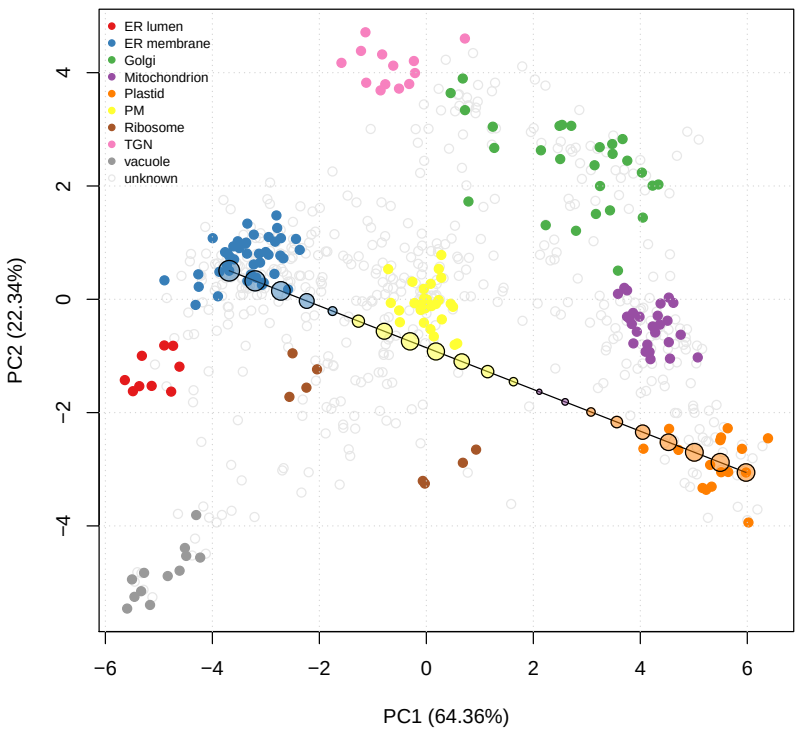
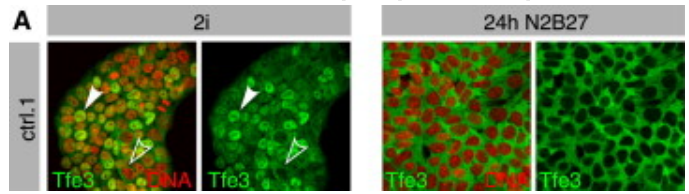


Figure : V-ATPase subunit d1 (P51683) with uncertain localisation between the endosome and lysosome.

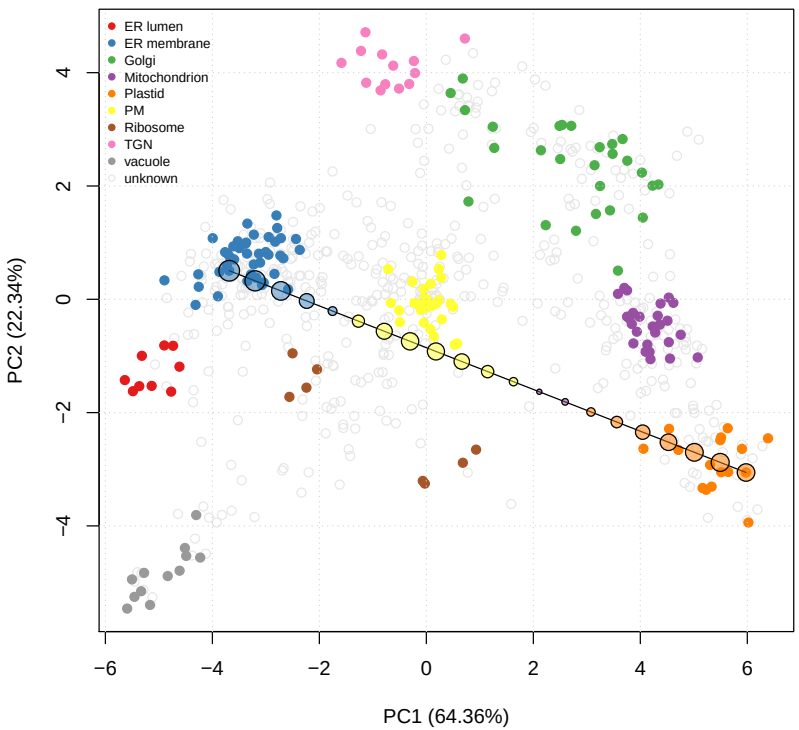
Dual-localisation Proteins may be present simultaneously in several organelles (e.g. trafficking). Simulation on *A. thaliana* data from Dunkley et al. (2006) (Gatto et al., 2014b) (left). Example from embryonic stem cells (Christoforou et al., 2016) (right).

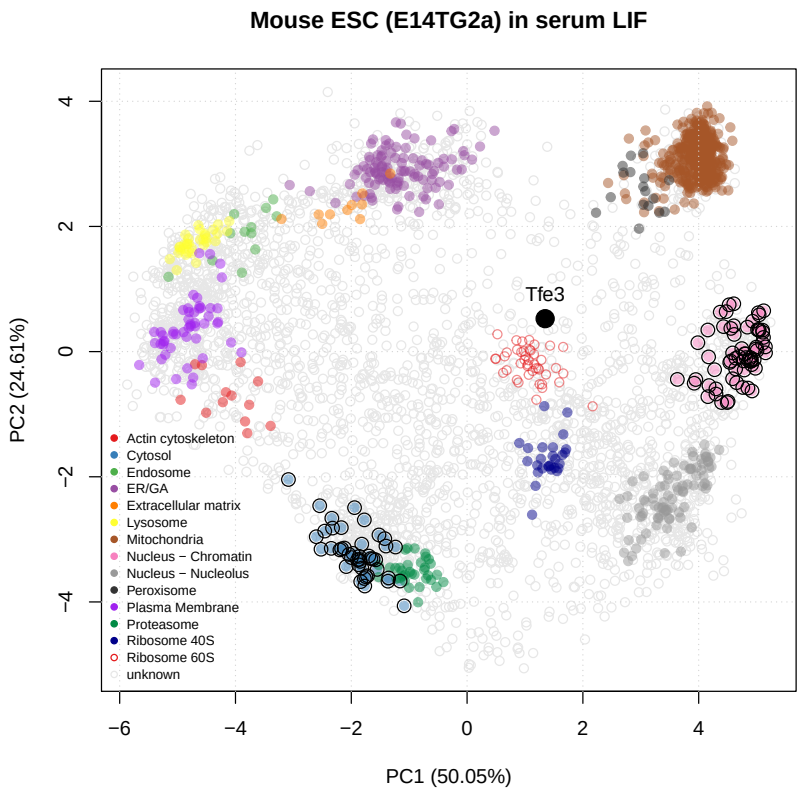


Dual-localisation Proteins may be present simultaneously in several organelles (e.g. trafficking). Simulation on *A. thaliana* data from Dunkley et al. (2006) (Gatto et al., 2014b) (left). Example from embryonic stem cells (Christoforou et al., 2016) (right).



From Betschinger et al. (2013)





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** at the Cambridge Centre for Proteomics.

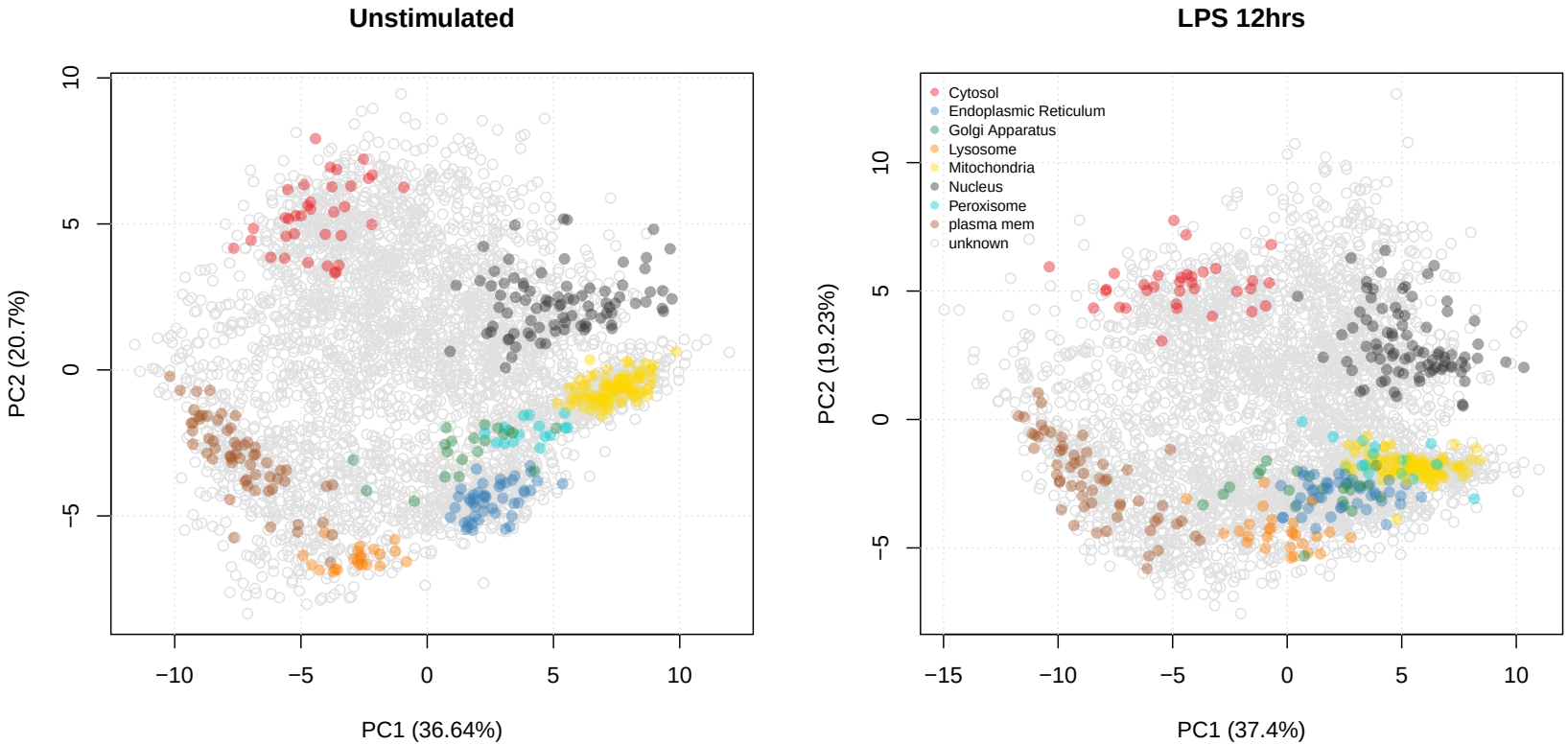


Figure : Spatial maps of unstimulated and LPS-treated cells (combined triplicates).

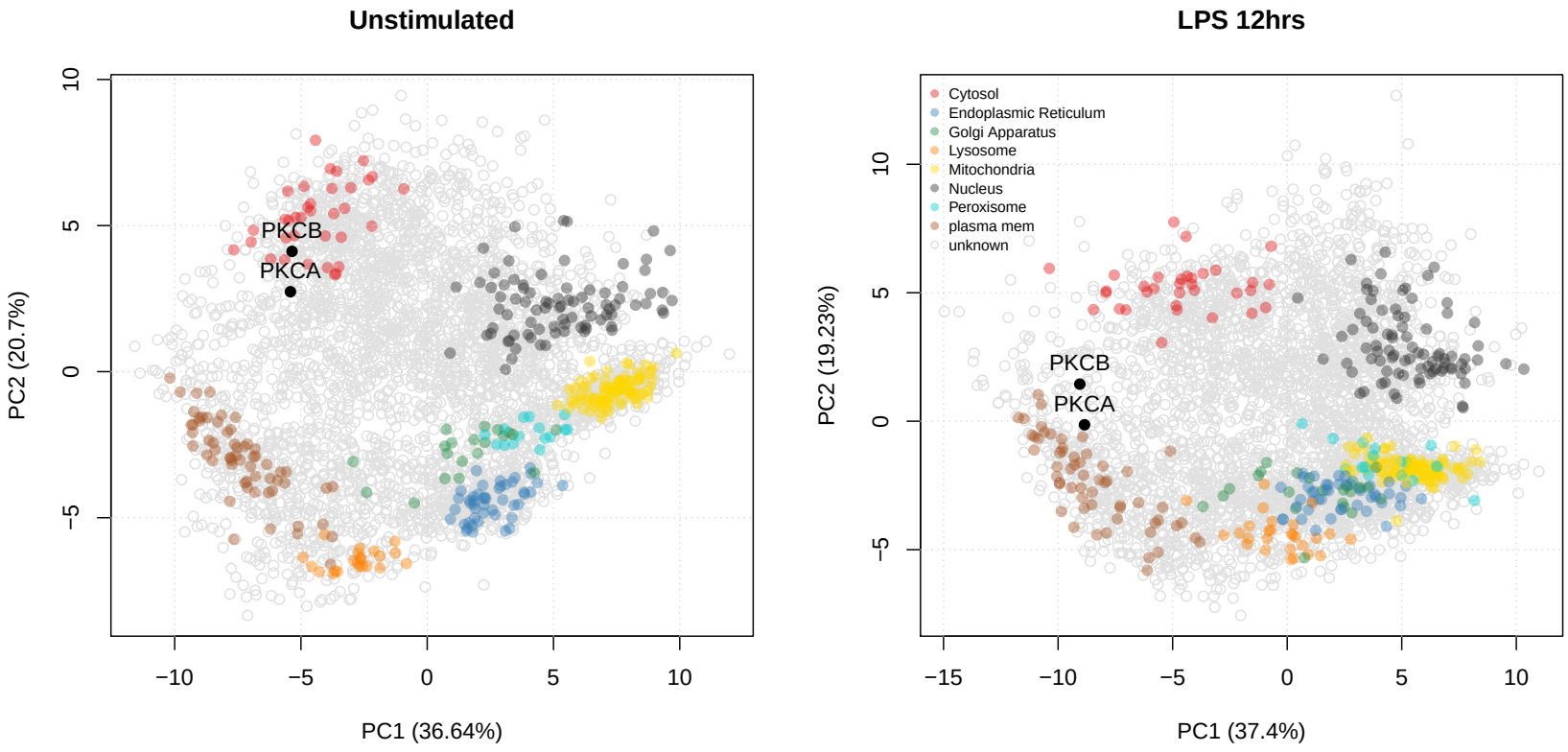


Figure : Relocation of Protein Kinase C α and β from the cytosol to the plasma membrane, **driving maturation into a differentiated macrophage phenotype**.

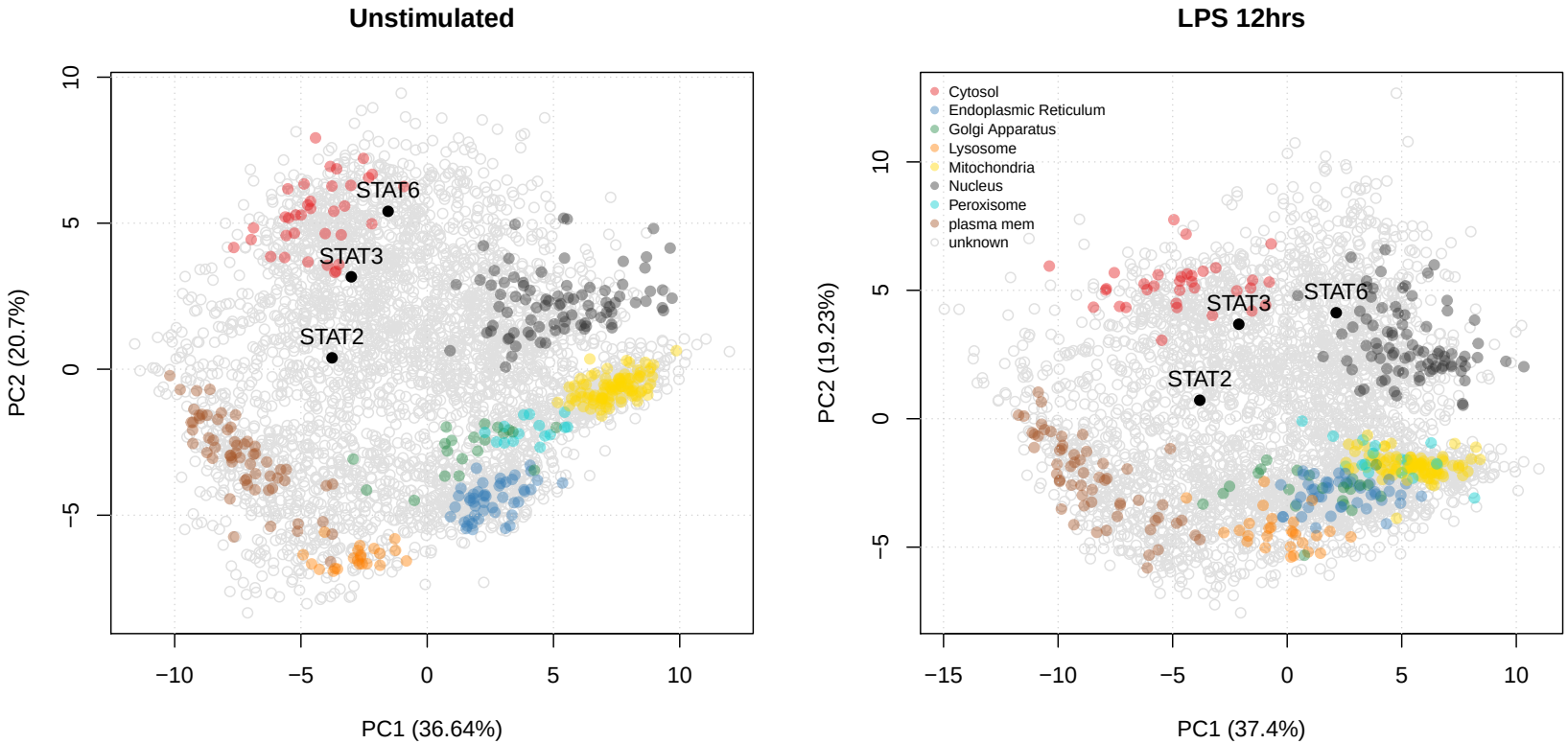


Figure : Relocation of Signal transducer and activator of transcription 6 (STAT6) from the cytosol to the Nucleus, **activating anti-bacterial and anti-viral-like response**. Validated by microscopy and see also Chen et al. (2011).

Computational infrastructure

Reliance on computational biology to acquire reliable biological knowledge.

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**.

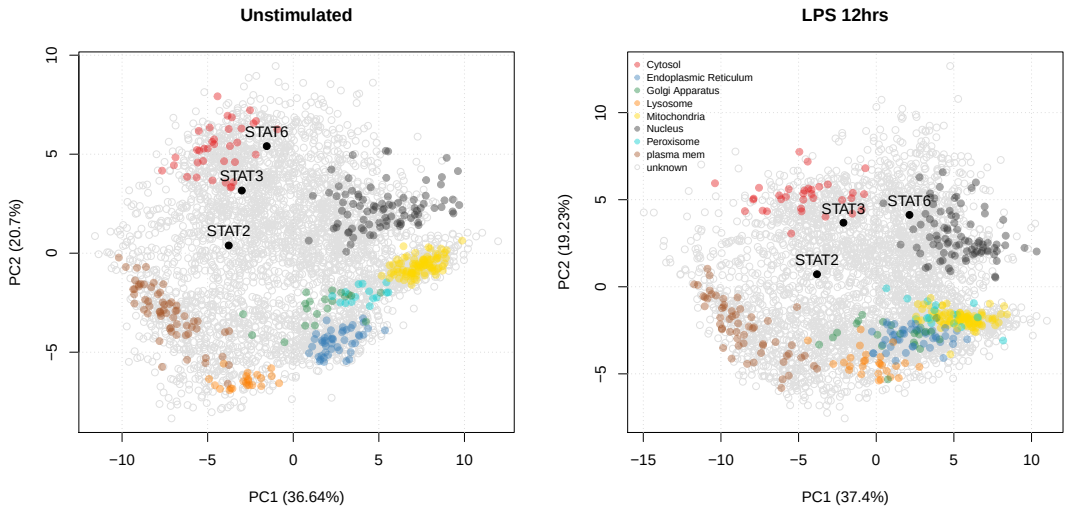
[1] ... which are all reproducible, by the way.

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

[3] between and within domains/software

Conclusions

1. Protein sub-cellular localisation: technologies (hyperLOPIT) and opportunities (sub-cellular maps, multi- and trans-localisation).



2. Reliance on computational biology and dedicated software to interpret data and acquire biological knowledge.

```
> library("pRoloc")
```

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.

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.

References II

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.

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.

P Wu and TG Dietterich. Improving svm accuracy by training on auxiliary data sources. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, New York, NY, USA, 2004. ACM.

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/1180393

Thank you for your attention

Supplementary slides: Computational infrastructure

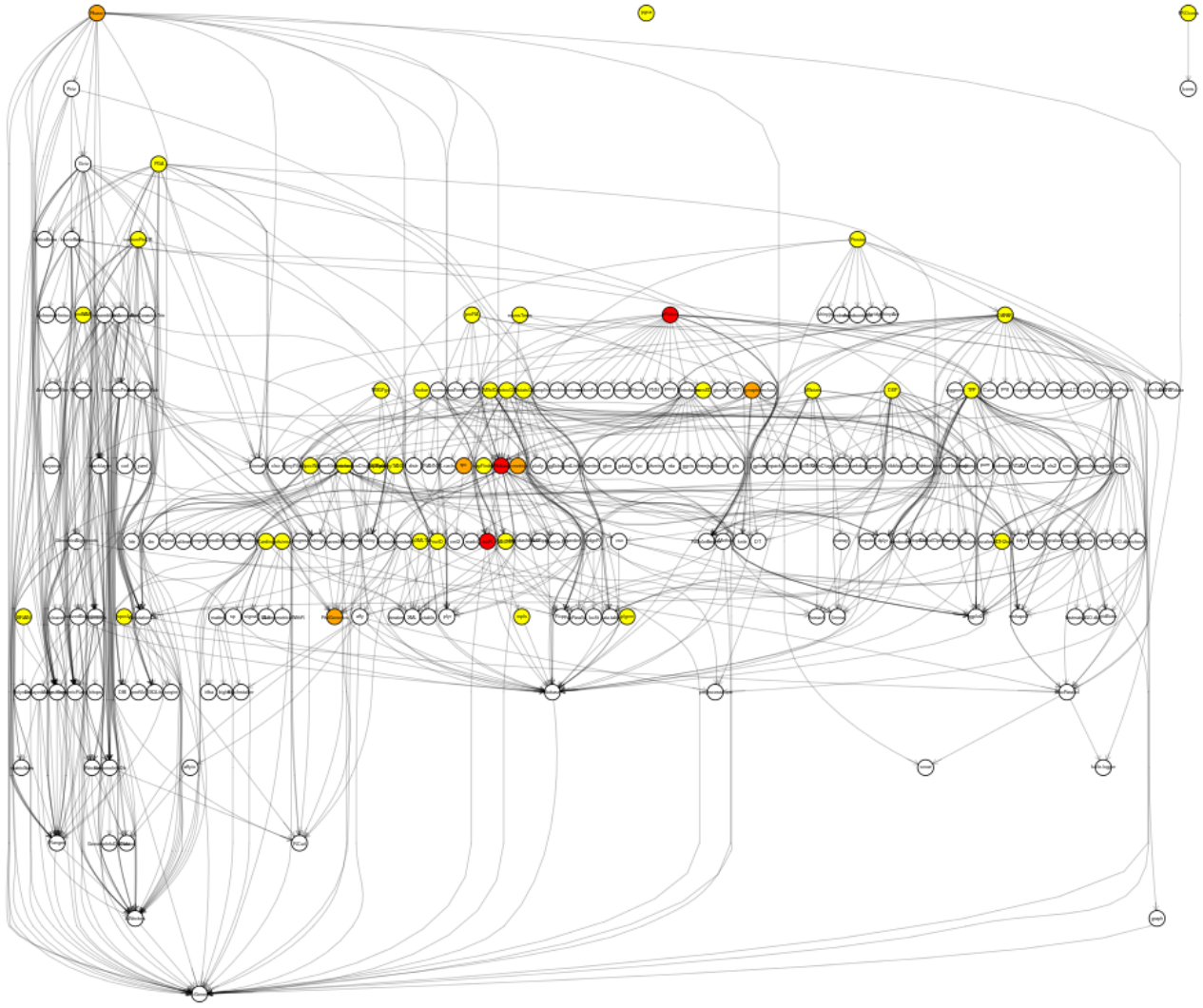


Figure : **Collaboration between packages**: Dependency graph containing 41 MS and proteomics-tagged packages (out of 100+) and their dependencies.

MSnbase example

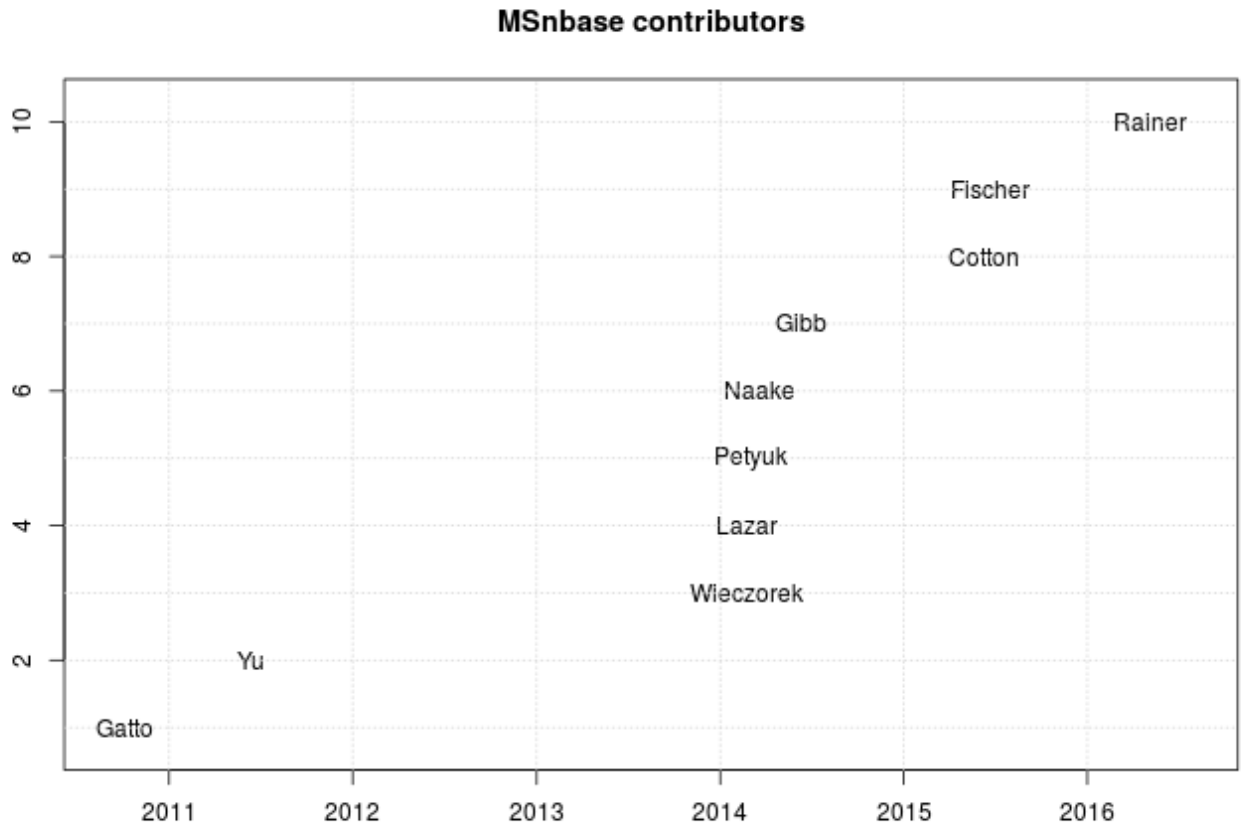


Figure : **Collaboration within packages**: Contributions to the `MSnbase` package (1220 downloads from unique IP addresses in January 2018) since its creation, the last one leading to **common proteomics/metabolomics infrastructure**. More details: https://lgatto.github.io/msnbase-contribs/

Supplementary slides: tranfer learning

Application to PPI/Protein complexes

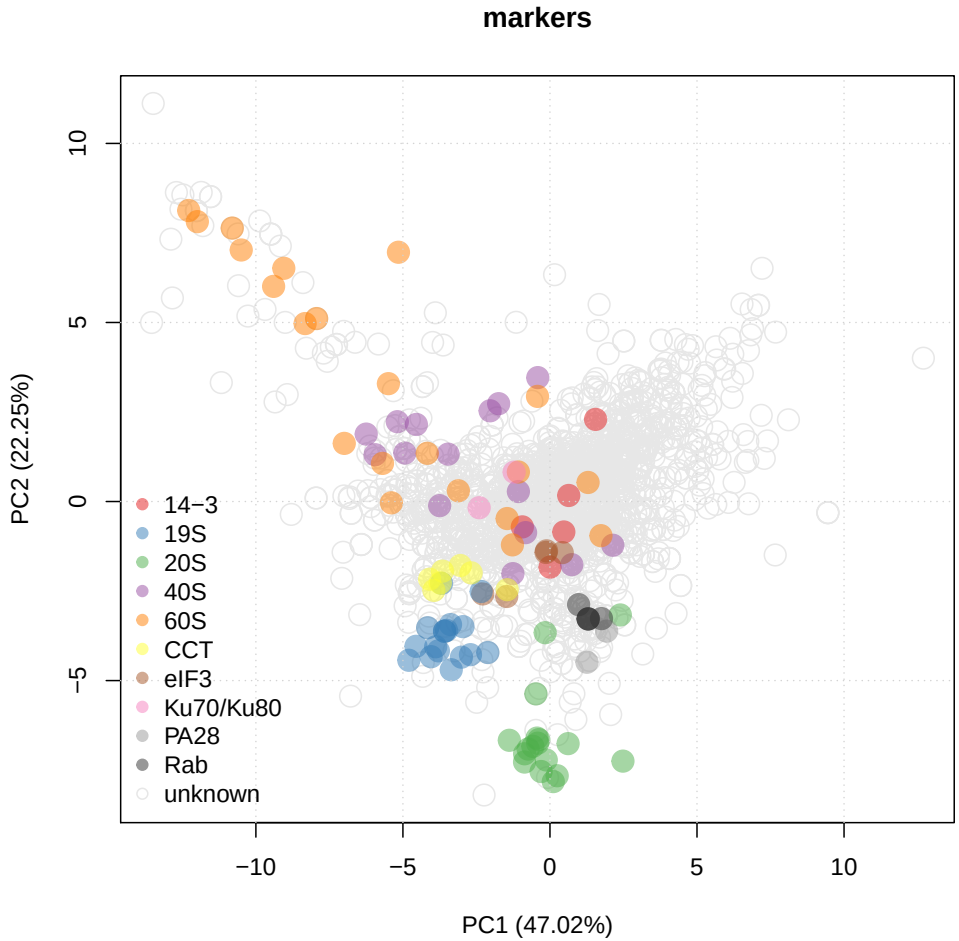
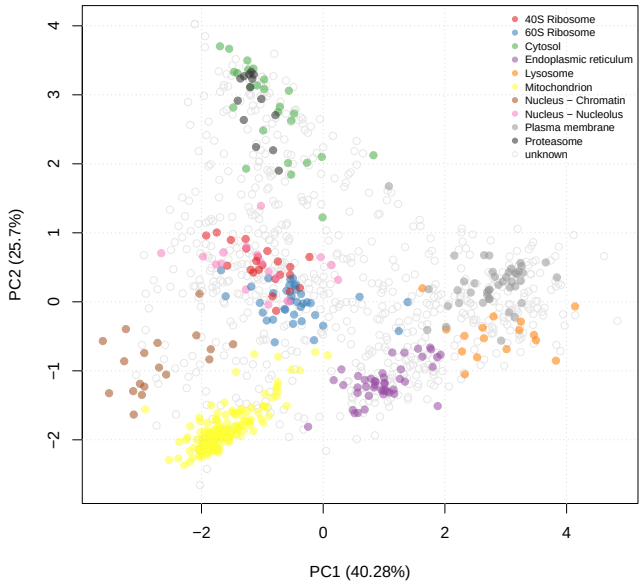


Figure : Data on proteasome complexes from Fabre *et al.* Mol Syst Biol (2015), DOI: 10.15252/msb.20145497

Transfer learnig, based on Wu and Dietterich (2004):

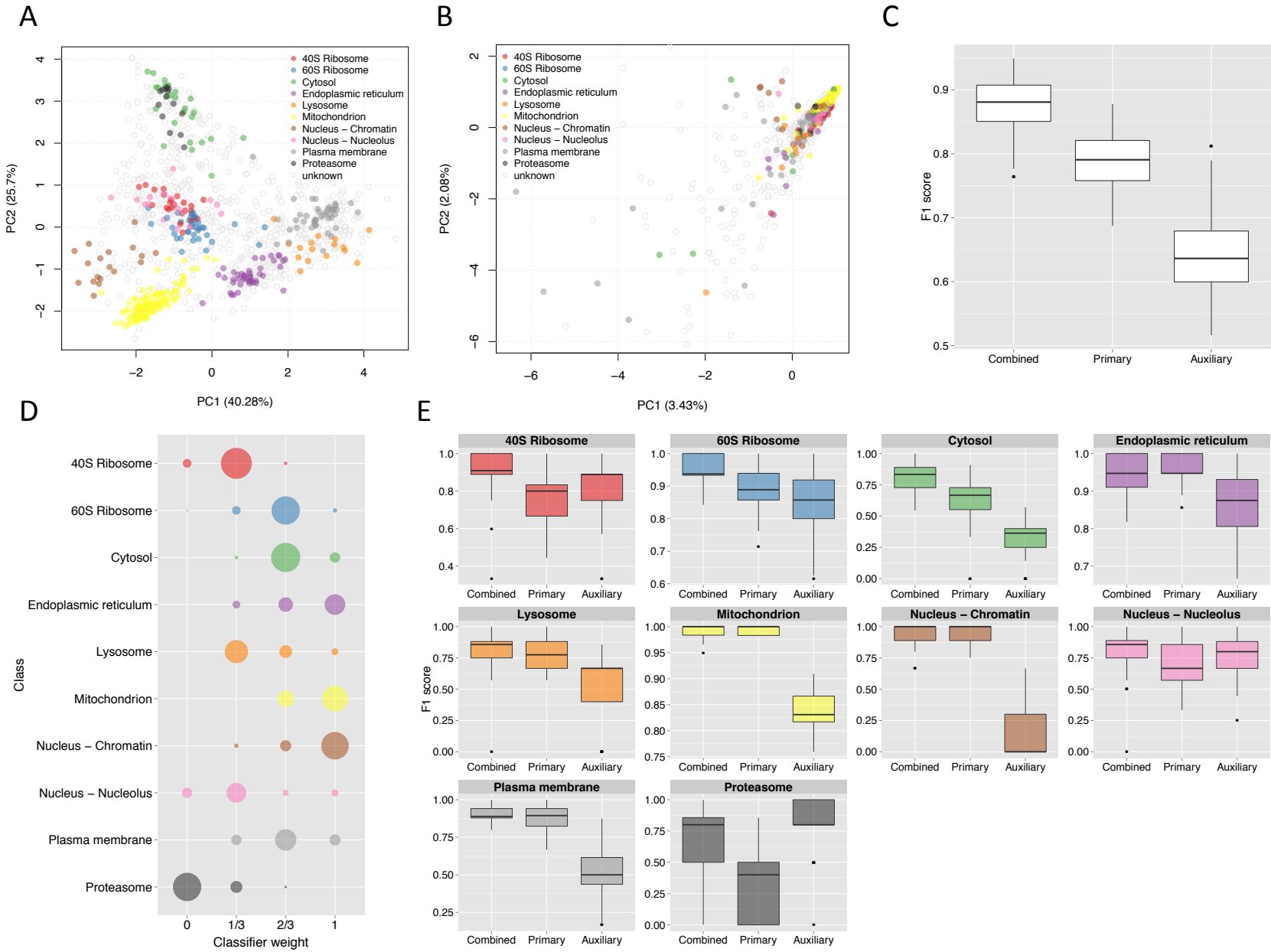
Class-weighted kNN

$$V(c_i)_j = \theta^* n_{ij}^P + (1 - \theta^*) n_{ij}^A$$



Linear programming SVM

$$f(\mathbf{x}, \mathbf{v}; \boldsymbol{\alpha}_P, \boldsymbol{\alpha}_A, b) = \sum_{l=1}^{m} y_l \left[\alpha_l^P K^P(\mathbf{x}_l, \mathbf{x}) + \alpha_l^A K^A(\mathbf{v}_l, \mathbf{v}) \right] + b$$



Data from mouse stem cells (E14TG2a).

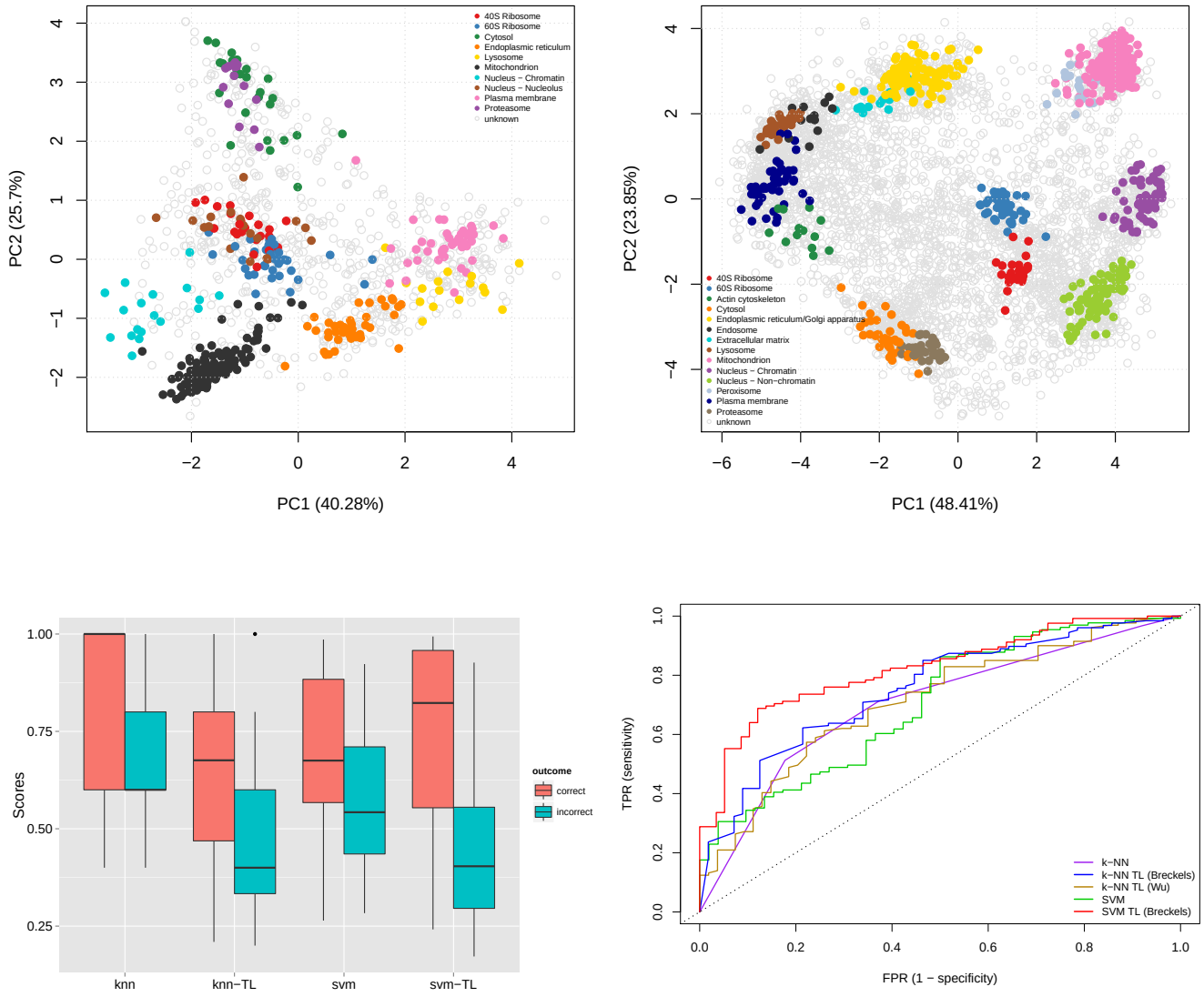


Figure : From Breckels et al. (2016a) *Learning from heterogeneous data sources: an application in spatial proteomics*.