

Mapping the sub-cellular proteome

Laurent Gatto

lg390@cam.ac.uk – @lgatt0

<http://cpu.sysbiol.cam.ac.uk/>

Slides @ <http://bit.ly/20171108FLI>

DOI [10.5281/zenodo.1042726](https://doi.org/10.5281/zenodo.1042726)

Last update: November 6, 2017

8 Nov 2017, Fritz Lipmann Institute

Plan

Spatial proteomics

- The LOPIT pipeline

- Improving on LOPIT

 - Experimental advances: hyperLOPIT

 - Computational advances: Transfer learning

- Biological applications

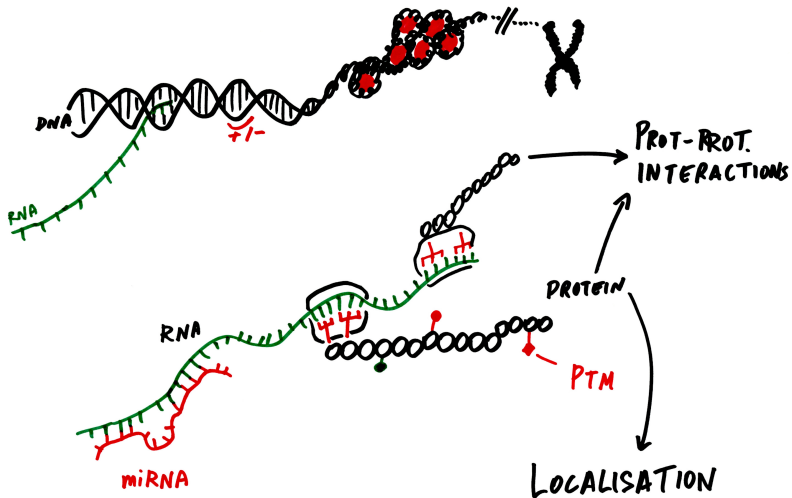
 - Dual-localisation

 - Trans-localisation

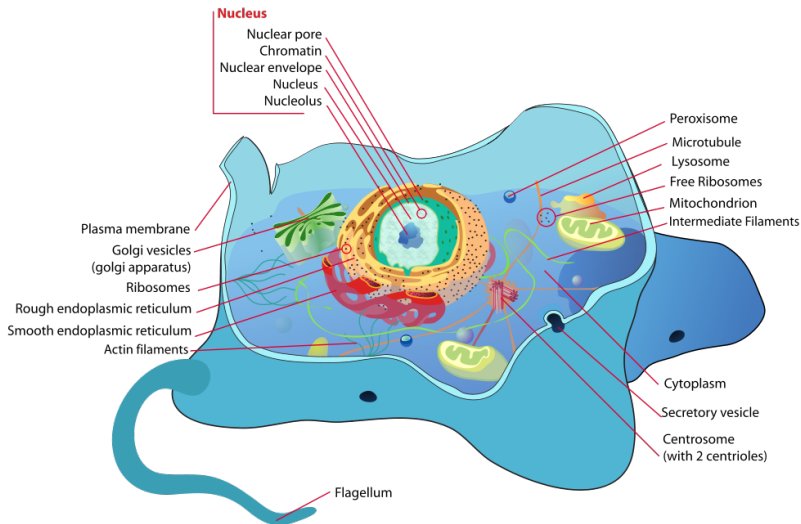
- R/Bioconductor software

- Open development

Regulations



Cell organisation



Spatial proteomics is the systematic study of protein localisations.

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.

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).

Re-localisation in

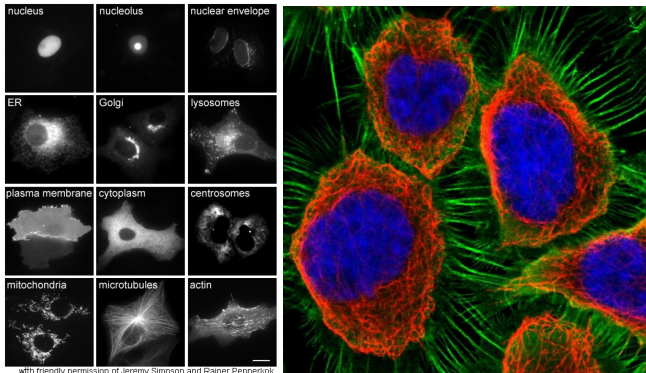
- ▶ **Differentiation**: Tfe3 in mouse ESC (Betschinger et al., 2013).
- ▶ **Metabolism**: changes in carbon sources, elemental limitations.

Spatial proteomics - How, experimentally

Single cell direct observation	Population level				
	Subcellular fractionation (number of fractions)				
	1 fraction	2 fractions (enriched and crude)	n discrete fractions	n continuous fractions (gradient approaches)	
	GFP Epitope Prot.-spec. antibody	Pure fraction catalogue	Subtractive proteomics (enrichment)	Invariant rich fraction (clustering)	PCP (χ^2)
Cataloguing		Relative abundance			
Tagging	Quantitative mass spectrometry				

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

Fusion proteins and immunofluorescence



Fusion proteins and immunofluorescence

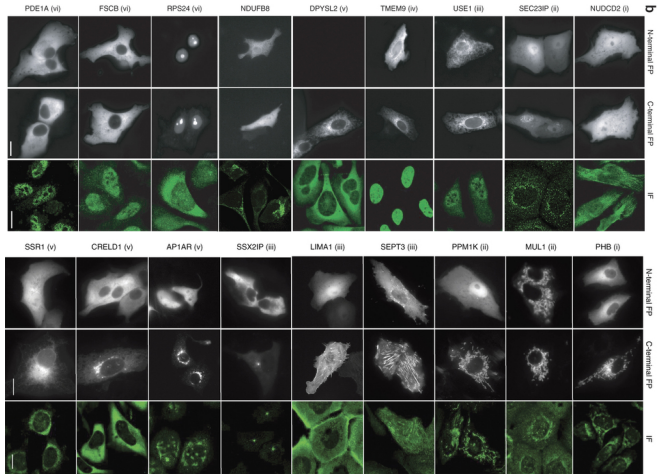


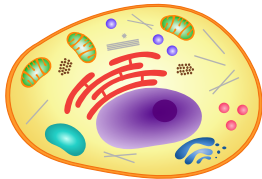
Figure : Example of discrepancies between IF and FPs as well as between FP tagging at the N and C termini (Stadler et al., 2013).

Spatial proteomics - How, experimentally

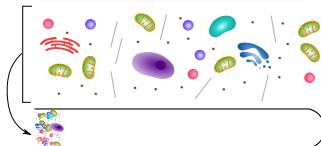
Single cell direct observation	Population level				
	Subcellular fractionation (number of fractions)				
	1 fraction	2 fractions (enriched and crude)	n discrete fractions	n continuous fractions (gradient approaches)	
	GFP Epitope Prot.-spec. antibody	Pure fraction catalogue	Subtractive proteomics (enrichment)	Invariant rich fraction (clustering)	PCP (χ^2)
Cataloguing		Relative abundance			
Tagging	Quantitative mass spectrometry				

Figure : Organelle proteomics approaches (Gatto et al., 2010). Gradient approaches: Dunkley et al. (2006), Foster et al. (2006).

⇒ **Explorative/discovery** approaches, **global localisation maps**.

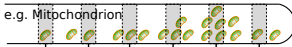


Cell lysis



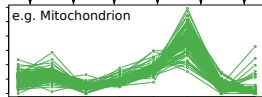
Fractionation/centrifugation

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}	<i>loc₁</i>
p ₃	q _{3,1}	q _{3,2}	...	q _{3, m}	unknown
p ₄	q _{4,1}	q _{4,2}	...	q _{4, m}	<i>loc_i</i>
⋮	⋮	⋮	⋮	⋮	⋮
p _j	q _{j,1}	q _{j,2}	...	q _{j, m}	unknown

Visualisation and classification

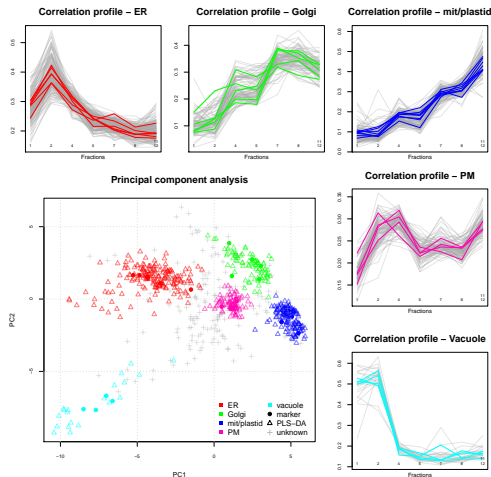
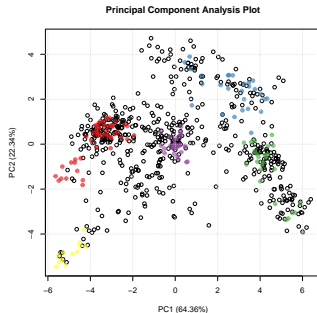


Figure : From Gatto et al. (2010), *Arabidopsis thaliana* data from Dunkley et al. (2006)

Data analysis

	Fraction ₁	Fraction ₂	...	Fraction _m		markers	
prot ₁	q _{1,1}	q _{1,2}	...	q _{1, m}	...	unknown	...
prot ₂	q _{2,1}	q _{2,2}	...	q _{2, m}		organelle ₁	
prot ₃	q _{3,1}	q _{3,2}	...	q _{3, m}		unknown	
prot ₄	q _{4,1}	q _{4,2}	...	q _{4, m}		organelle ₂	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
prot _i	q _{i,1}	q _{i,2}	...	q _{i, m}		⋮	
⋮	⋮	⋮	⋮	⋮	⋮	organelle _k	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
prot _n	q _{n,1}	q _{n,2}	...	q _{n, m}	...	unknown	...
	Fraction ₁	Fraction ₂	...	Fraction _m			
			
	⋮	⋮	⋮	⋮			
			



Supervised machine learning

Using labelled marker proteins to match unlabelled proteins (of unknown localisation) with similar profiles and classify them as residents to the markers organelle class.

Current approaches - supervised ML

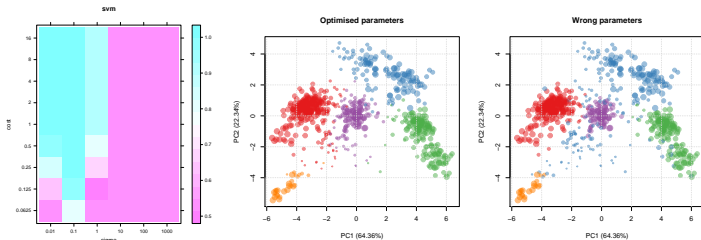
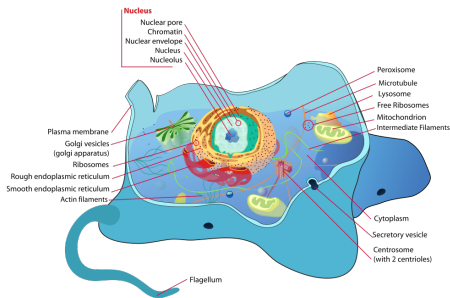
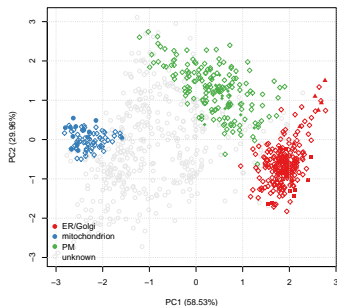


Figure : Support vector machines classifier with a radial basis kernel function, using the **pRoloc** Bioconductor package (Gatto et al., 2014a).

Limitations



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

Novelty detection

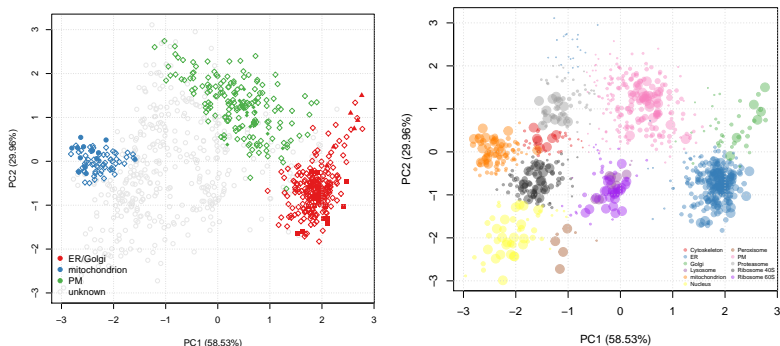


Figure : Left: *Drosophila* data from Tan et al. (2009). Right: Semi-supervised learning, Breckels et al. (2013).

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.

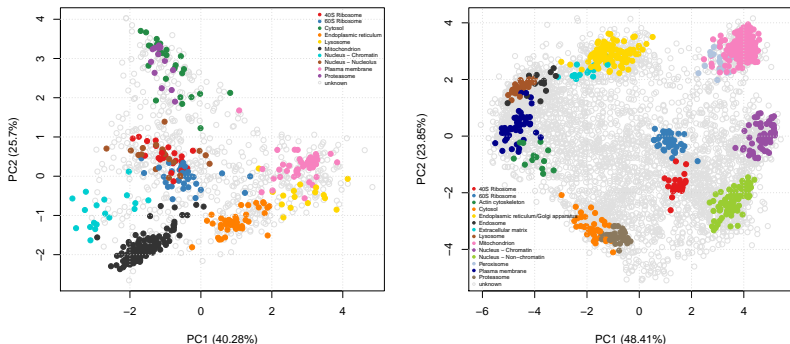


Figure : E14TG2a embryonic stem cells: old (left) vs. new, better resolved (right) experiments (Christoforou et al. (2016)).

Improving on LOPIT

LOPIT Dunkley et al. (2006)	Computational: <i>transfer learning</i> Breckels et al. (2016)
Experimental: <i>hyperLOPIT</i> Christoforou et al. (2016) Mulvey et al. (2017)	Biological discoveries

hyperLOPIT

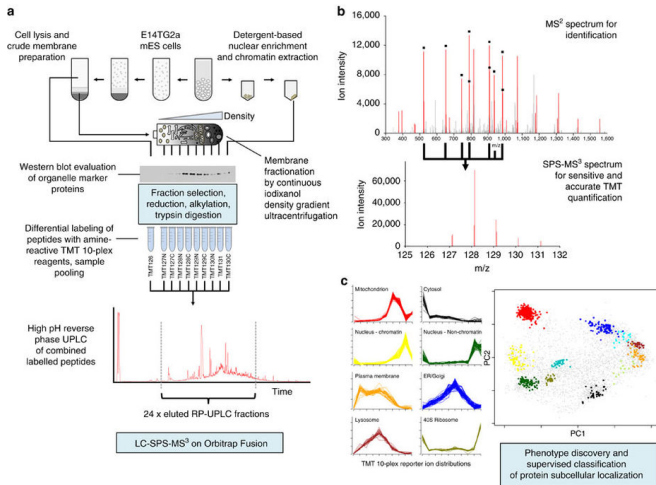


Figure : From Christoforou et al. (2016) *A draft map of the mouse pluripotent stem cell spatial proteome.*

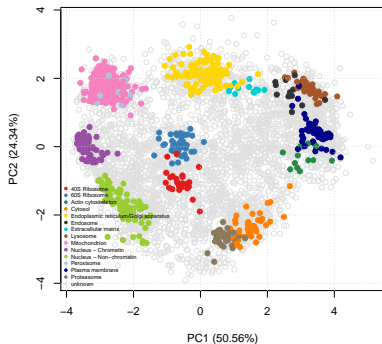
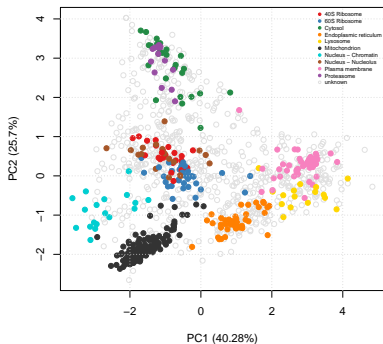


Figure : LOPIT on 8 fractions (using iTRAQ 8-plex) and 1109 proteins vs. hyperLOPIT on 10 fractions (using TMT 10-plex) and SPS-MS³ for 5032 proteins.

Transfer learning

What about annotation data from repositories such as the Gene Ontology (GO), sequence features, signal peptide, transmembrane domains, images, prediction software, ...

- ▶ From a user perspective: "**free/cheap**" vs. expensive
- ▶ 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**

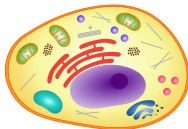
Transfer learning

What about annotation data from repositories such as the Gene Ontology (GO), sequence features, signal peptide, transmembrane domains, images, prediction software, ...

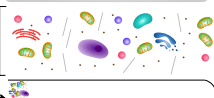
Transfer learning

Support/complement the **primary** target domain (experimental data) with **auxiliary** data (annotation, imaging, PPI, ...) features without compromising the integrity of our primary data (Breckels et al., 2016).

PRIMARY EXPERIMENTAL DATA



Cell lysis

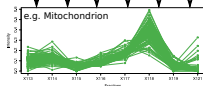


Fractionation/centrifugation

e.g. Mitochondrion

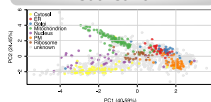


Quantitation/identification by mass spectrometry



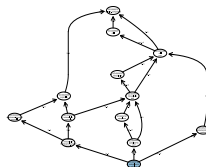
	X110	X114	X115	X116	X117	X118	X119	X121
CD327F7	0.1362	0.1350	0.1062	0.1487	0.2777	0.1429	0.0380	0.0010
PF14486	0.1014	0.1020	0.0946	0.1061	0.1207	0.0996	0.0180	0.0077
CD327A3	0.1297	0.1201	0.0946	0.1061	0.2760	0.1463	0.0206	0.0060
CD327C5	0.1008	0.1007	0.0919	0.1061	0.1461	0.0900	0.0000	0.0000

Visualisation



Database query

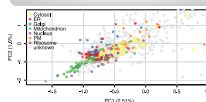
Extract GO CC terms



Convert terms to binary

	GO:0005822	GO:0005739	GO:0005783	GO:
GO:0005822	1	1	1	...
GO:0005739	1	1	1	...
GO:0005783	1	1	1	...
GO:0005822	1	1	1	...
GO:0005739	1	1	1	...
GO:0005783	1	1	1	...

Visualisation

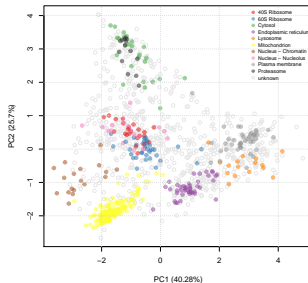


AUXILIARY DRY DATA

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

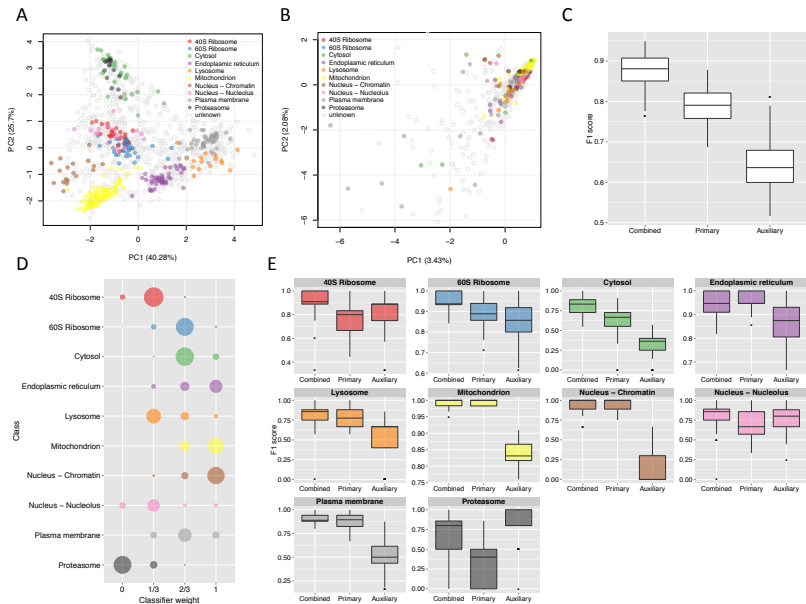
Class-weighted kNN

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



Linear programming SVM

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



Data from mouse stem cells (E14TG2a).

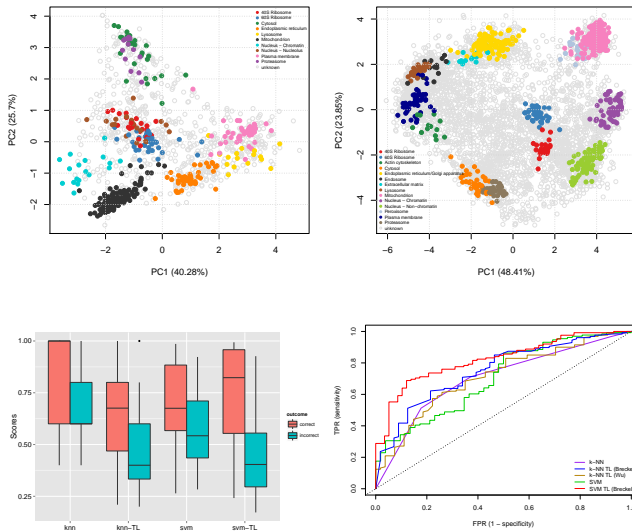


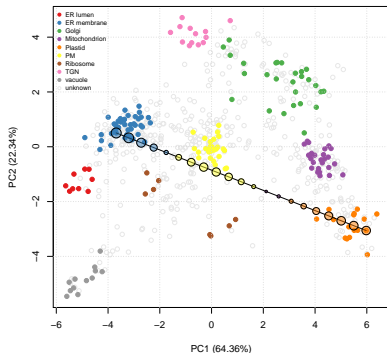
Figure : From Breckels et al. (2016) *Learning from heterogeneous data sources: an application in spatial proteomics.*

Biological applications

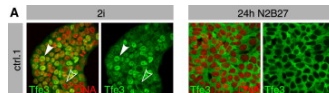
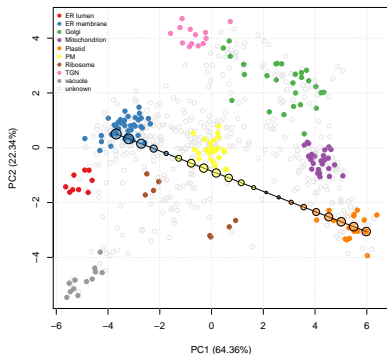
- ▶ Multi-localisation
- ▶ Trans-localisation

Dependent on good sub-cellular resolution.

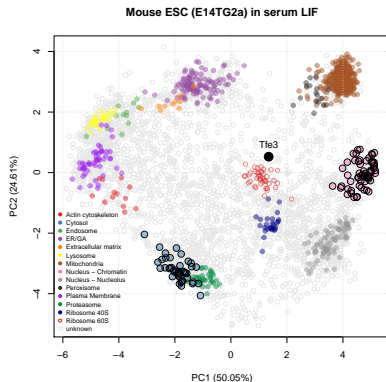
Dual-localisation Proteins may be present simultaneously in several organelles (e.g. trafficking).



Dual-localisation Proteins may be present simultaneously in several organelles (e.g. trafficking).

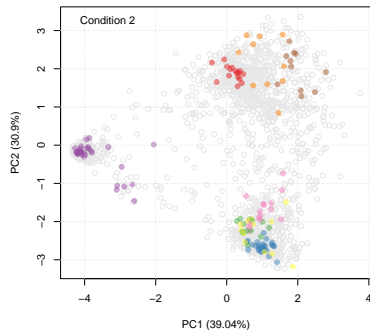
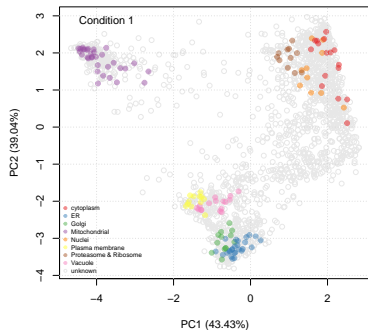


From Betschinger et al. (2013)



Spatial dynamics

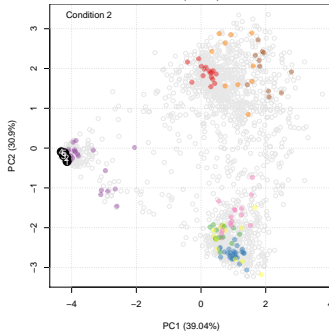
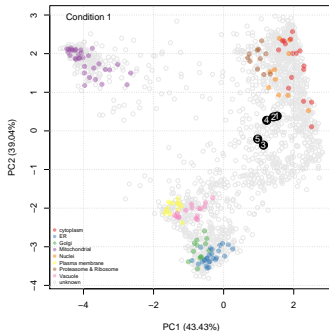
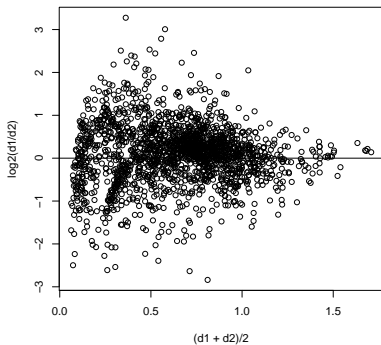
Trans-localisation Changes in localisation upon perturbations.



Spatial dynamics

$$d_1 = \text{dist}(\text{profile}_{\text{condition}_1}^{\text{rep}_1}, \text{profile}_{\text{condition}_2}^{\text{rep}_1})$$

$$d_2 = \text{dist}(\text{profile}_{\text{condition}_1}^{\text{rep}_2}, \text{profile}_{\text{condition}_2}^{\text{rep}_2})$$



Beyond organelles: application to PPI/Protein complexes

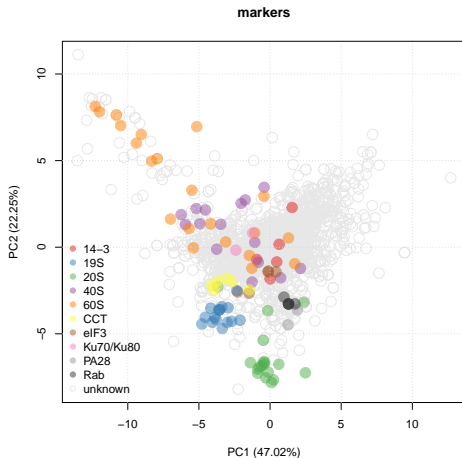


Figure : Data on proteasome complexes from Fabre *et al.* Mol Syst Biol (2015), DOI: [10.15252/msb.20145497](https://doi.org/10.15252/msb.20145497)

Plan

Spatial proteomics

- The LOPIT pipeline

- Improving on LOPIT

 - Experimental advances: hyperLOPIT

 - Computational advances: Transfer learning

- Biological applications

 - Dual-localisation

 - Trans-localisation

R/Bioconductor software

Open development

R/Bioconductor:

- ▶ Software for spatial proteomics.
- ▶ Ecosystem for high throughput biology data analysis and comprehension.

Software for mass spectrometry and (spatial) proteomics

Bioconductor Open source, enable **reproducible research**, enables understanding of the data (not a black box) and **drive scientific innovation**.

- ▶ **MSnbase** – infrastructure to handle quantitative data and meta-data (Gatto and Lilley, 2012) (~500 unique IP download/month in 2016).
- ▶ **pRoloc** and **pRolocGUI** – dedicated visualisation and ML infrastructure for spatial proteomics (Gatto et al., 2014a) (~200 unique IP download/month in 2016).
- ▶ **pRolocdata** – structured and annotated spatial proteomics data (Gatto et al., 2014a).
- ▶ And more generally **RforProteomics** (Gatto and Christoforou, 2014) (~160 unique IP download/month in 2016).

Plan

Spatial proteomics

- The LOPIT pipeline

- Improving on LOPIT

 - Experimental advances: hyperLOPIT

 - Computational advances: Transfer learning

- Biological applications

 - Dual-localisation

 - Trans-localisation

R/Bioconductor software

Open development

- ▶ What is Collaborative and open development?
- ▶ Use case: MSnbase and mzR: contributors and shared infrastructure for MS-based proteomics and metabolomics.

Use case

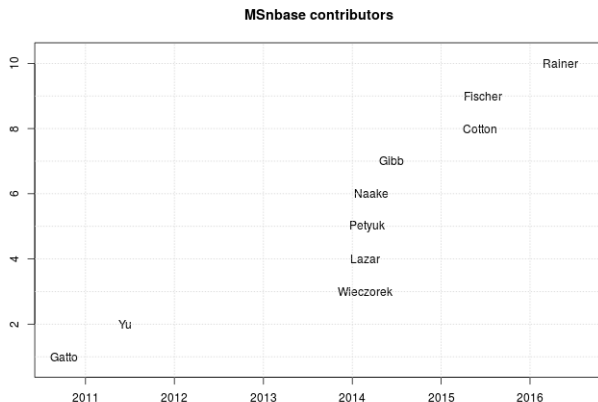


Figure : Contributions to the MSnbase package since its creation. More details: <https://lgatto.github.io/msnbase-contribs/>

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 2016. doi: 10.1371/journal.pcbi.1004920.
- 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.
- 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.


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 A Christoforou. Using R and Bioconductor for proteomics data analysis. *Biochim Biophys Acta*, 1844(1 Pt A):42–51, Jan 2014.
- 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 Wiczorek, T Burger, and K S Lilley. Mass-spectrometry based spatial proteomics data analysis using pRoloc and pRolocdata. *Bioinformatics*, Jan 2014a.

References III

- 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.
- 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 Britov?ek, 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.
- 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

- ▶ **Lisa Breckels**, Computational Proteomics Unit, Cambridge (ML, algo)
- ▶ **Kathryn Lilley**, Cambridge Centre of Proteomics (Proteomics)
- ▶ **Funding**: BBSRC, Wellcome Trust
- ▶ Slides: <http://bit.ly/20171108FLI> DOI [10.5281/zenodo.1042726](https://doi.org/10.5281/zenodo.1042726)
- ▶ License: 

Thank you for your attention