Mapping the sub-cellular proteome

Laurent Gatto
lg390@cam.ac.uk – @lgatt0
http://www.damtp.cam.ac.uk/user/lg390/

Slides @ https://zenodo.org/record/1063508
DOI 10.5281/zenodo.1063508

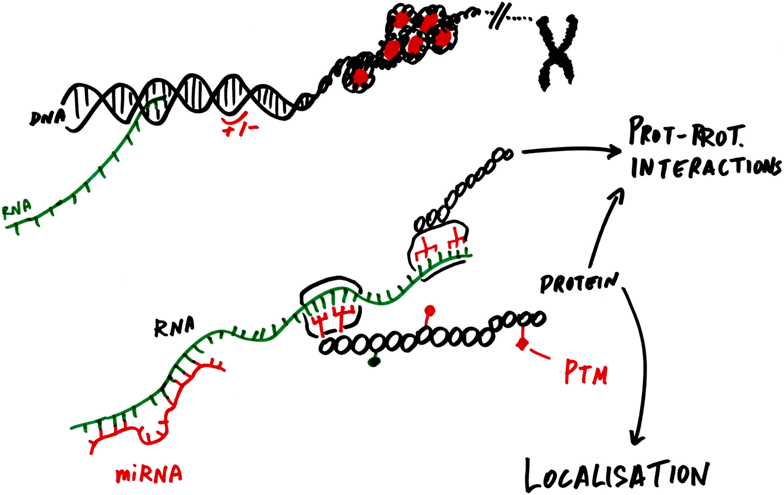
22 Nov 2017, Cambridge Computational Biology Institute

Plan

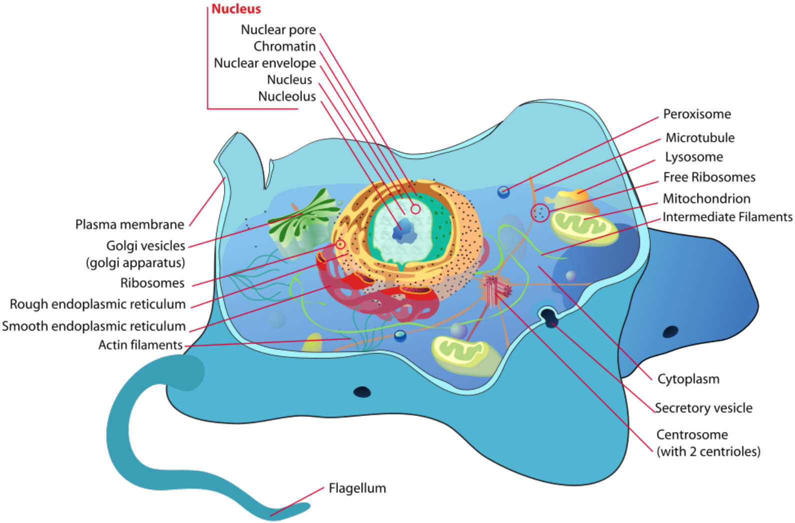
- Spatial proteomics
  - The LOPIT pipeline
  - Improving on LOPIT
    - Experimental advances: hyperLOPIT
    - Computational advances: Transfer learning
  - Biological applications
    - Dual-localisation
    - Trans-localisation

- Open development: R/Bioconductor software

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: Tfe3 in mouse ESC (Betschinger et al., 2013).
- Activation of biological processes.

Examples later.

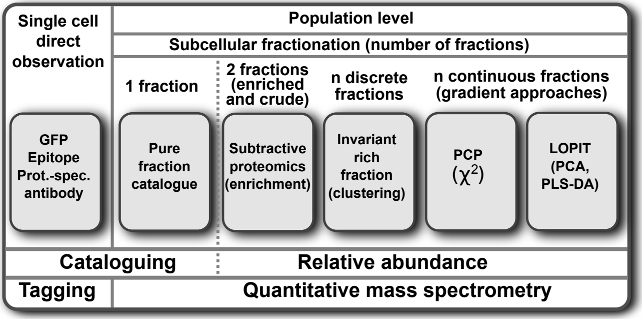
Spatial proteomics - Why?

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).

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)

LOPIT (PCA, PLS-DA)

Cataloguing

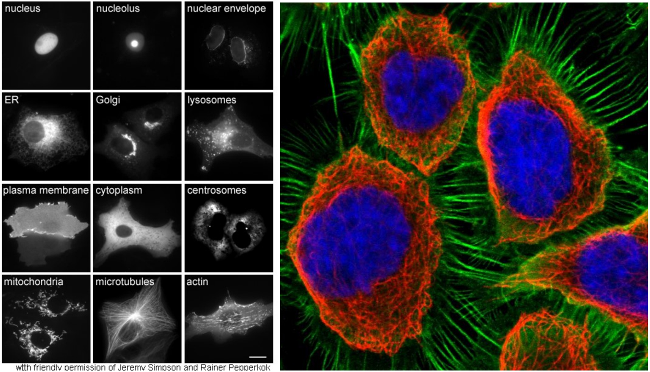
Relative abundance

Tagging

Quantitative mass spectrometry

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

Fusion proteins and immunofluorescence



with friendly permission of Jeremy Simpson and Rainer Pepperkok

Figure : Targeted protein localisation.

Fusion proteins and immunofluorescence

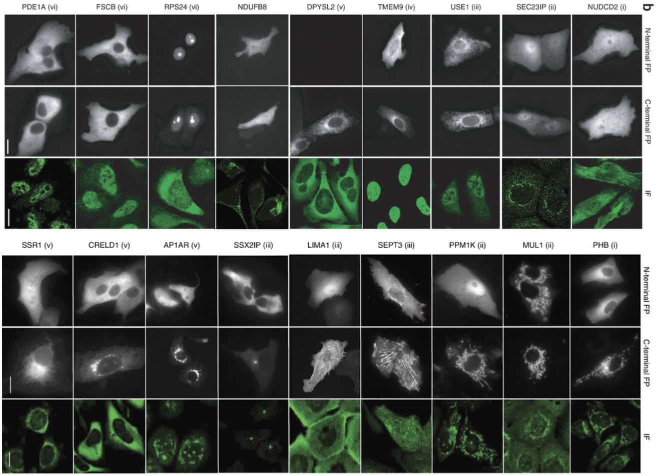
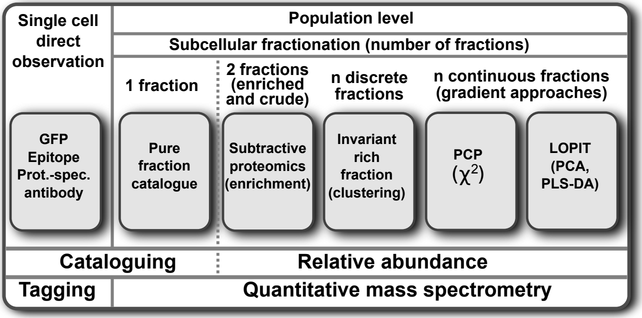


Figure : Example of discrepancies between IF and FPs as well as between FP tagging at the N and C termini (Stadler et al., 2013).

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)

LOPIT (PCA, PLS-DA)

Cataloguing

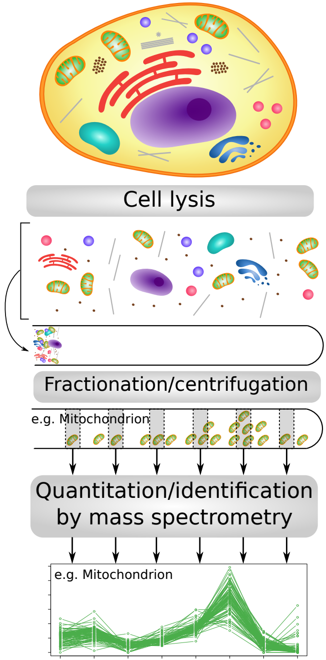
Relative abundance

Tagging

Quantitative mass spectrometry

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

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



Cell lysis

Fractionation/centrifugation

Quantitation/identification by mass spectrometry

Quantitation data and organelle markers

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

Visualisation and classification

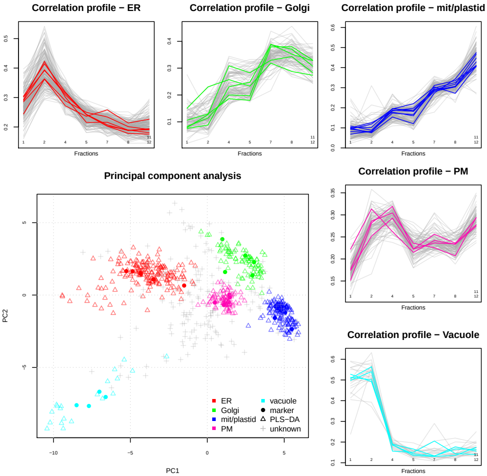
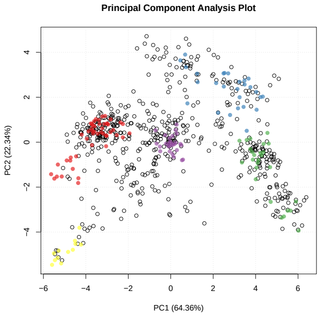


Figure : From Gatto et al. (2010), *Arabidopsis thaliana* data from Dunkley et al. (2006)

Data analysis

	Fraction_1	Fraction_2	…	Fraction_m		markers	
prot_1	$q_{1,1}$	$q_{1,2}$	…	$q_{1,\,m}$	…	*unknown*	…
prot_2	$q_{2,1}$	$q_{2,2}$	…	$q_{2,\,m}$		$\mathit{organelle}_1$	
prot_3	$q_{3,1}$	$q_{3,2}$	…	$q_{3,\,m}$		*unknown*	
prot_4	$q_{4,1}$	$q_{4,2}$	…	$q_{4,\,m}$		$\mathit{organelle}_2$	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
prot_i	$q_{i,1}$	$q_{i,2}$	…	$q_{i,\,m}$		$\mathit{organelle}_k$	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
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.

Supervised ML

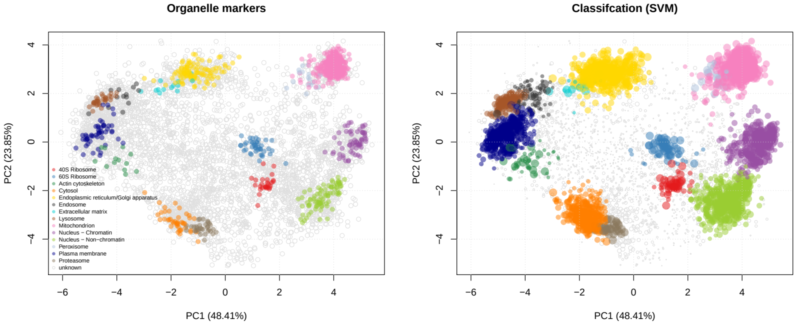
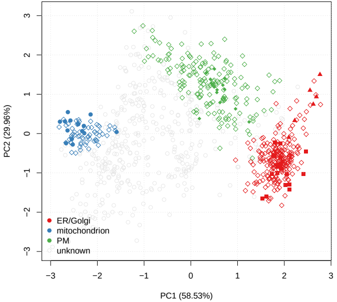
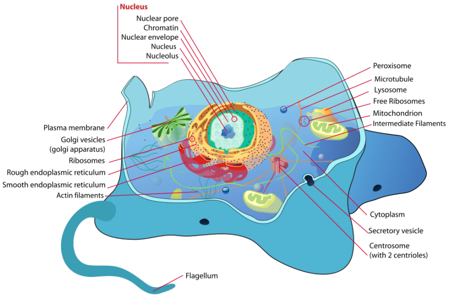


Figure : Support vector machines classifier (after classification cutoff) on the embryonic stem cell data from Christoforou et al. (2016).

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

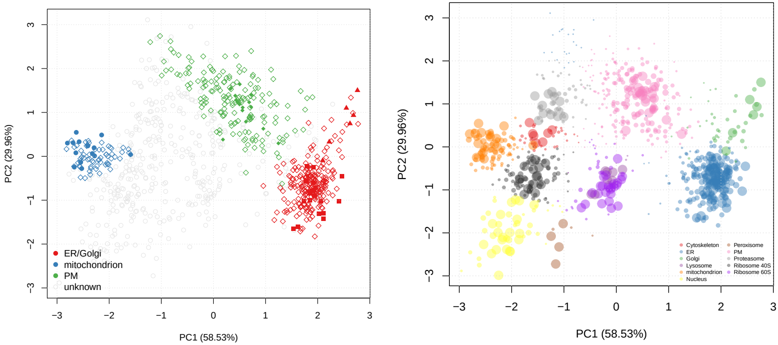


Figure : Left: Original *Drosophila* data from Tan et al. (2009). Right: After semi-supervised learning and classification, Breckels et al. (2013).

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.

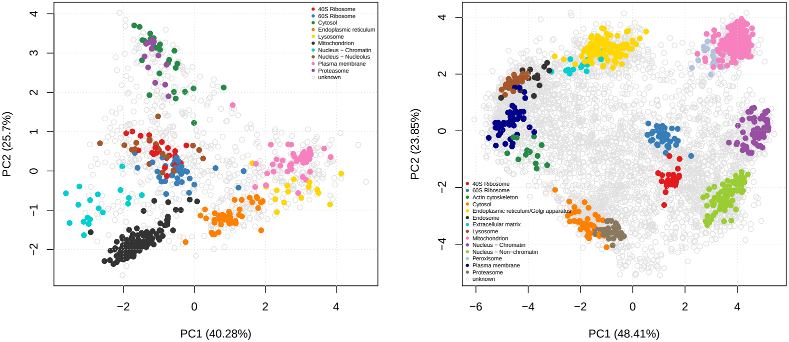


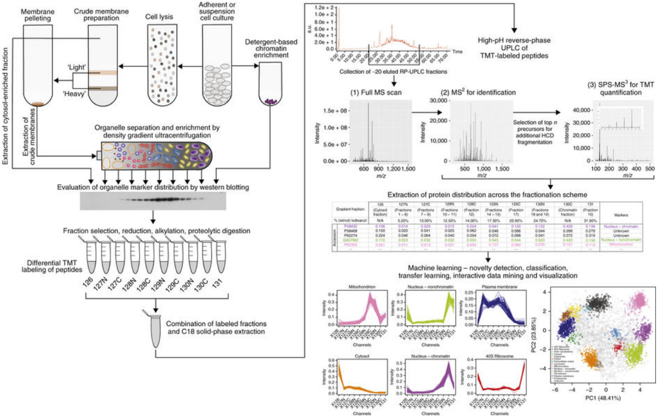
Figure : E14TG2a embryonic stem cells: old (left) *vs.* new, better resolved (right) experiments (Christoforou et al. (2016)).

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 ⇒ **biological discoveries**.

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


Membrane pelleting
Crude membrane preparation
Cell lysis
Adherent or suspension cell culture
Detergent-based chromatin enrichment
'Light'
'Heavy'
Extraction of cytosol-enriched fraction
Extraction of crude membranes
Organelle separation and enrichment by density gradient ultracentrifugation
Evaluation of organelle marker distribution by western blotting
Fraction selection, reduction, alkylation, proteolytic digestion
Differential TMT labeling of peptides
126
127N
127C
128N
128C
129N
129C
130N
130C
131
Combination of labeled fractions and C18 solid-phase extraction
a.u.
Time
Collection of ~20 eluted RP-UPLC fractions
High-pH reverse-phase UPLC of TMT-labeled peptides
(1) Full MS scan
(2) MS² for identification
(3) SPS-MS³ for TMT quantification
Selection of top n precursors for additional HCD fragmentation
Intensity
m/z
Extraction of protein distribution across the fractionation scheme
Machine learning – novelty detection, classification, transfer learning, interactive data mining and visualization
Mitochondrion
Nucleus – nonchromatin
Plasma membrane
Cytosol
Nucleus – chromatin
40S Ribosome
Channels
PC1 (48.41%)
PC2 (23.85%)

Gradient fraction:	126 (Cytosol fraction)	127N (Fractions 1 – 6)	127C (Fractions 7 – 9)	128N (Fractions 10 – 11)	128C (Fraction 12)	129N (Fractions 14 – 15)	129C (Fraction 17)	130N (Fractions 18 and 19)	130C (Chromatin fraction)	131 (Fraction 10)	Markers
% (wt/vol) Iodixanol	N/A	5.20%	10.00%	12.50%	14.00%	17.30%	20.90%	24.70%	N/A	31.90%	
Accession P48432	0.106	0.014	0.020	0.012	0.024	0.041	0.120	0.122	0.409	0.134	Nucleus – chromatin
P58468	0.150	0.020	0.041	0.025	0.062	0.046	0.066	0.044	0.266	0.279	Unknown
P62274	0.224	0.046	0.064	0.040	0.054	0.072	0.068	0.041	0.072	0.319	Unknown
Q9CRB2	0.172	0.023	0.032	0.032	0.055	0.045	0.044	0.020	0.420	0.156	Nucleus – nonchromatin
P62305	0.001	0.011	0.013	0.021	0.034	0.107	0.360	0.239	0.101	0.113	Mitochondrion
...	...	...	...	...	...	...	...	...	...	...	...

Figure : From Mulvey et al. (2017) *Using hyperLOPIT to perform high-resolution mapping of the spatial proteome*.

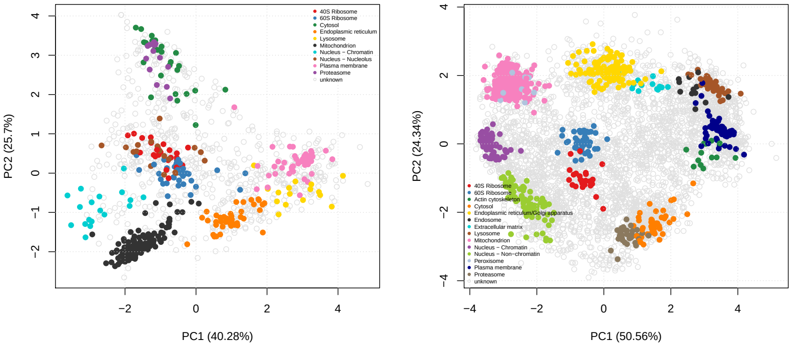


Figure : E14TG2a LOPIT on 8 fractions (using iTRAQ 8-plex) and 1109 proteins *vs.* hyperLOPIT on 10 fractions (using TMT 10-plex) and SPS-MS^3 for 5032 proteins.

Computational advances: Transfer learning

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

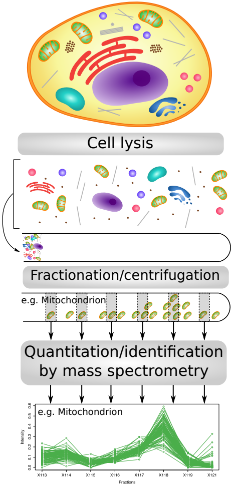
Transfer learning

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., 2016a).

PRIMARY EXPERIMENTAL DATA



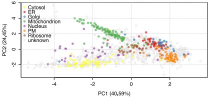
Cell lysis

Fractionation/centrifugation

Quantitation/identification by mass spectrometry

		X113	X114	X115	X116	X117	X118	X119	X121
x_1	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
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
x_n	.	.	.	.	.	.	.	.	.

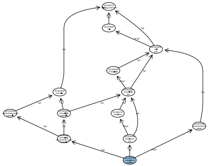
Visualisation



AUXILIARY DRY DATA

Database query

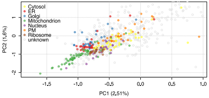
Extract GO CC terms



Convert terms to binary

		GO$_1$	...	...	...	...	GO$_A$
		GO:0016021	GO:0005789	GO:0005783	...	...	...
x_1	O00767	1	1	1	...	...	...
.	P51648	1	1	0	...	...	...
.	Q2TAA5	1	1	0	...	...	...
.	Q9UKV5	0	0	0	...	...	...
.	.	.	.	.			
.	.	.	.	.			
x_n	.	.	.	.			

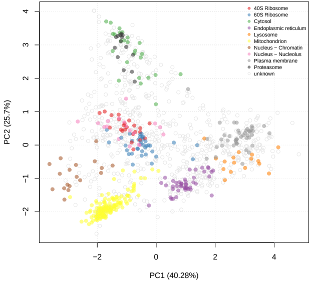
Visualisation



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

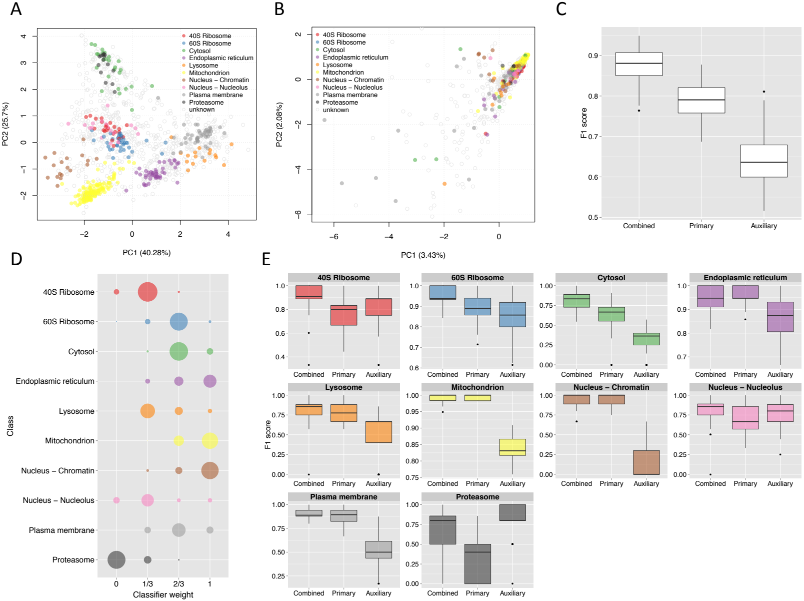
Class-weighted kNN

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



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).

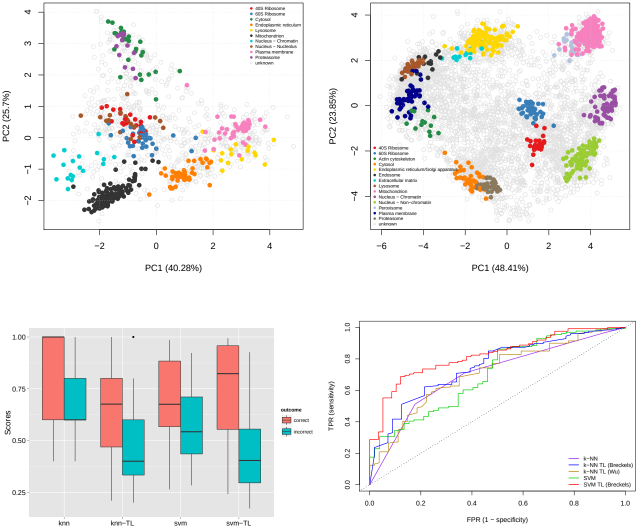


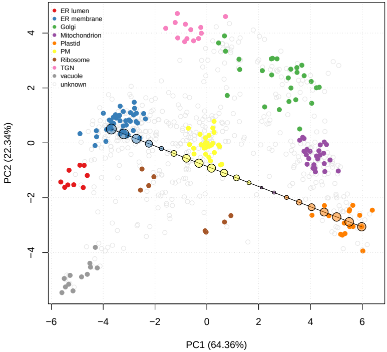
Figure : From Breckels et al. (2016a) *Learning from heterogeneous data sources: an application in spatial proteomics*.

Biological discoveries

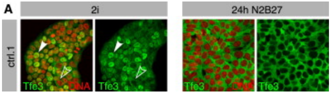
- Multi-localisation
- Trans-localisation

Dependent on good sub-cellular resolution.

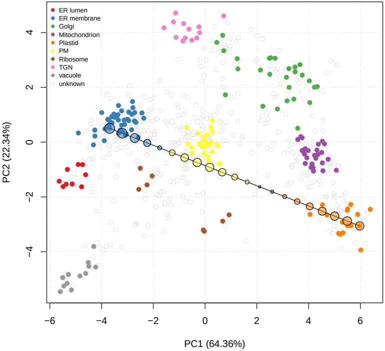
Dual-localisation Proteins may be present simultaneously in several organelles (e.g. trafficking). Simulation on *A. thaliana* data from Dunkley et al. (2006) (Gatto et al., 2014b) (left). Example from embryonic stem cells (Christoforou et al., 2016) (right).

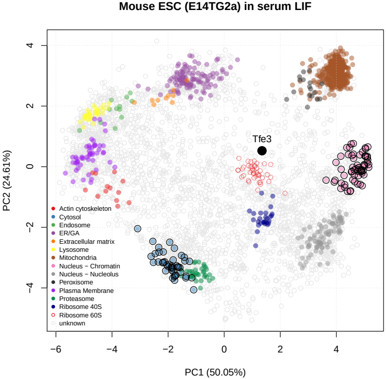


Dual-localisation Proteins may be present simultaneously in several organelles (e.g. trafficking). Simulation on *A. thaliana* data from Dunkley et al. (2006) (Gatto et al., 2014b) (left). Example from embryonic stem cells (Christoforou et al., 2016) (right).



From Betschinger et al. (2013)





Spatial dynamics

Trans-localisation event during monocyte to macrophage differenciation

Investigate the effect of 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 Centre for Proteomics.

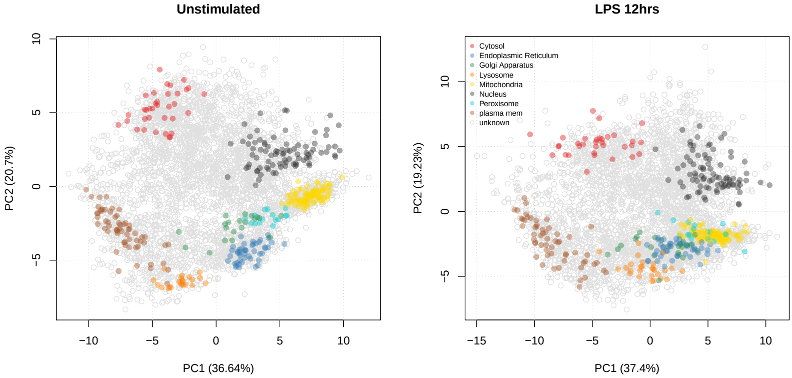


Figure : Spatial maps: unstimulated and LPS-treated.

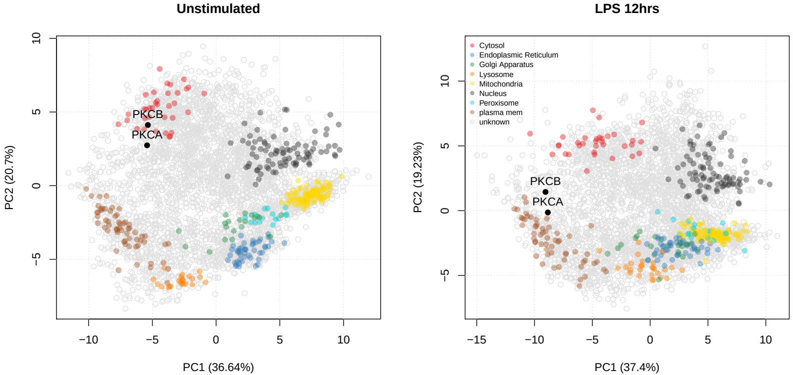


Figure : Relocation of Protein Kinase C alpha and beta from the cytosol to the plasma membrane, **driving maturation into a differentiated macrophage phenotype**.

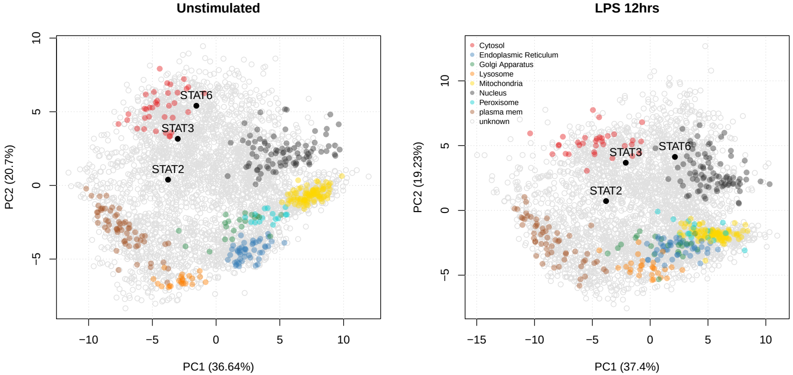


Figure : Relocation of Signal transducer and activator of transcription 6 (STAT6) from the cytosol to the Nucleus, **activating anti-bacterial and anti-viral-like response**. Validated by microscopy and see also Chen et al. (2011).

Beyond organelles: application to PPI/Protein complexes

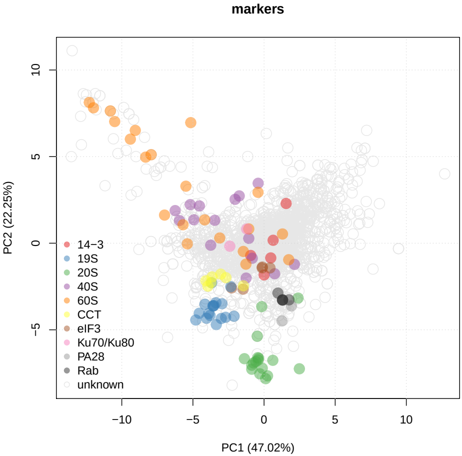


Figure : Data on proteasome complexes from Fabre *et al.* Mol Syst Biol (2015), DOI: 10.15252/msb.20145497

Plan

Spatial proteomics

The LOPIT pipeline

Improving on LOPIT

Experimental advances: hyperLOPIT

Computational advances: Transfer learning

Biological applications

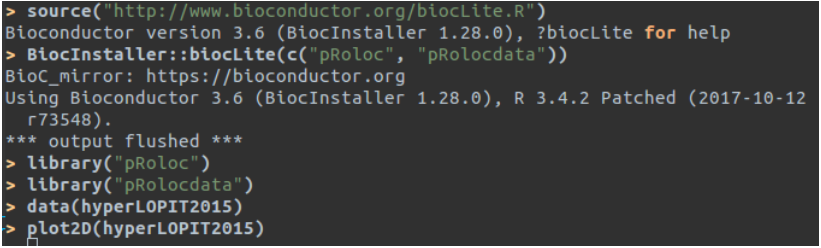
Dual-localisation

Trans-localisation

Open development: R/Bioconductor software

But none of this would matter if it wasn't **reproducible**!

Try it out yourselves:



```
> source("http://www.bioconductor.org/biocLite.R")
Bioconductor version 3.6 (BiocInstaller 1.28.0), ?biocLite for help
> BiocInstaller::biocLite(c("pRoloc", "pRolocdata"))
BioC_mirror: https://bioconductor.org
Using Bioconductor 3.6 (BiocInstaller 1.28.0), R 3.4.2 Patched (2017-10-12
  r73548).
*** output flushed ***
> library("pRoloc")
> library("pRolocdata")
> data(hyperLOPIT2015)
> plot2D(hyperLOPIT2015)
```

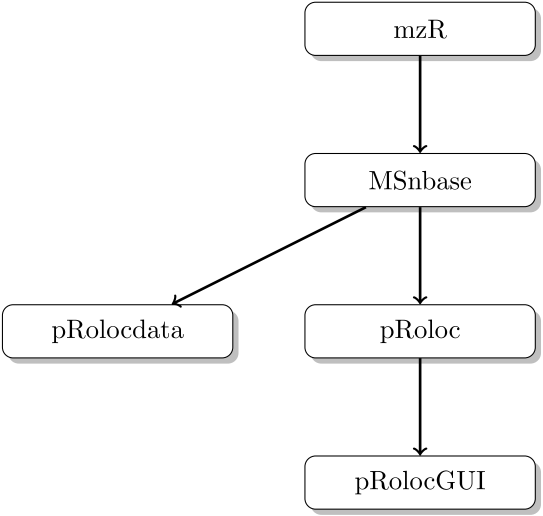
R/Bioconductor:

- 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**.

- `mzR` – low level access to raw and identification mass spectrometry data (Chambers and et al., 2012)
- `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). Try it out at `https://lgatto.shinyapps.io/christoforou2015/`
- `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).



mzR

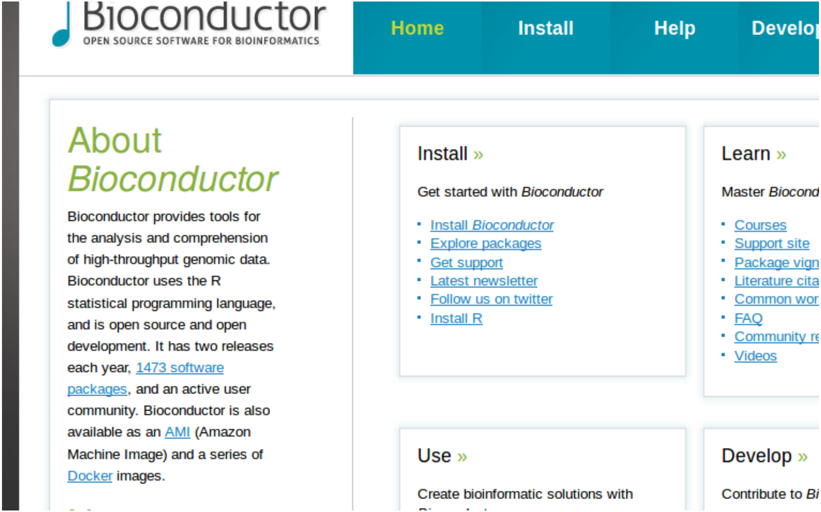
MSnbase

pRolocdata

pRoloc

pRolocGUI

http://www.bioconductor.org



About *Bioconductor*

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, 1473 software packages, and an active user community. Bioconductor is also available as an AMI (Amazon Machine Image) and a series of Docker images.

Install »

Get started with *Bioconductor*

- Install *Bioconductor*
- Explore packages
- Get support
- Latest newsletter
- Follow us on twitter
- Install R

Learn »

Master *Biocond*

- Courses
- Support site
- Package vign
- Literature cita
- Common wor
- FAQ
- Community re
- Videos

Use »

Create bioinformatic solutions with

Develop »

Contribute to *Bi*

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

- Bioconductor core team (lead by Dr. Martin Morgan)
- Common infrastructure
- Common documentation standards
- Common testing infrastructure
- Open package technical peer review

Quick getting started guide: https://lgatto.github.io/2017_11_09_Rcourse_Jena/navigating-the-bioconductor-project.html

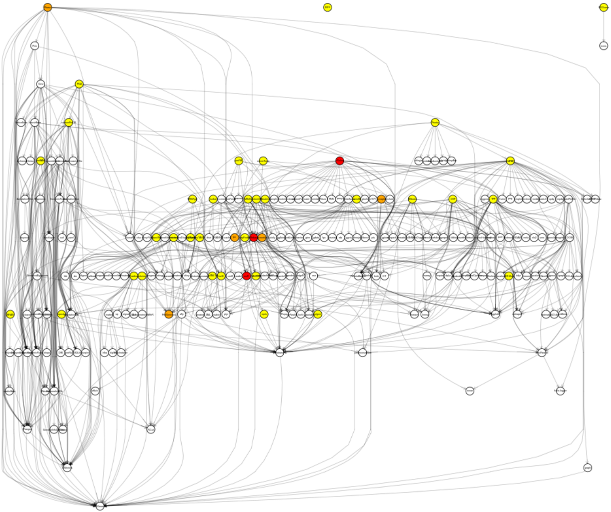


Figure : Dependency graph containing 41 MS and proteomics-tagged packages (out of 100+) and their dependencies. Showing all packages and deps would produce a big hairball.

MSnbase example

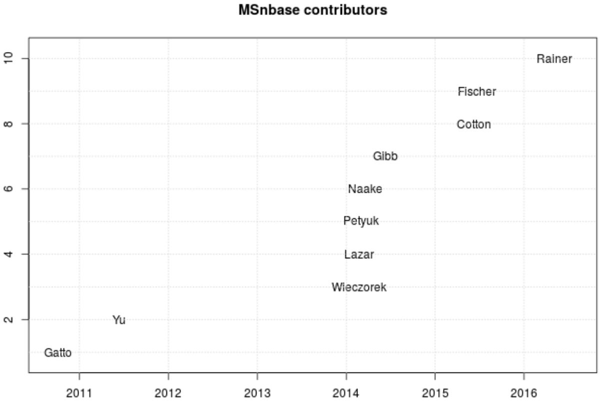


Figure : Contributions to the `MSnbase` package since its creation, the last one leading to **common proteomics/metabolomics infrastructure**. More details: `https://lgatto.github.io/msnbase-contribs/`

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 2016a. 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.

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.

MC Chambers and et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol*, 30(10): 918–20, Oct 2012.

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 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.

References II

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.

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: https://zenodo.org/record/1063508

Thank you for your attention