

Mapping the sub-cellular proteome

Laurent Gatto

`laurent.gatto@uclouvain.be`

de Duve Institute – UCLouvain

slides `http://bit.ly/20190322LIBST`

22 March 2019 – LIBST

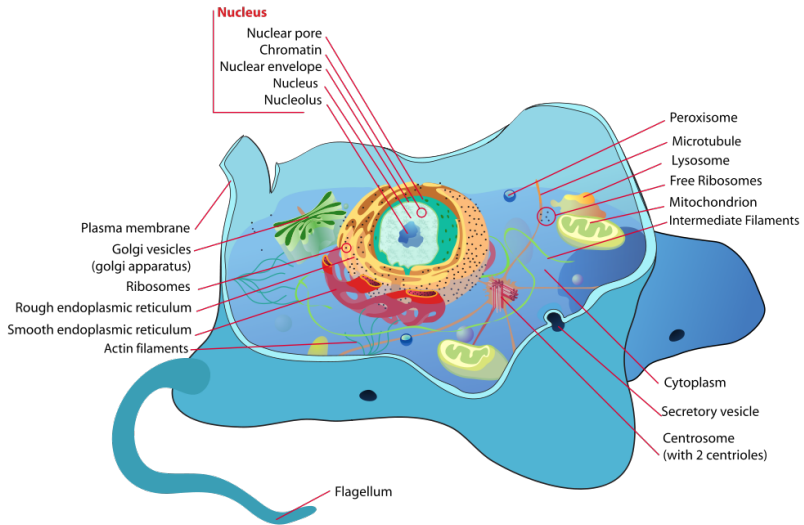
Abstract

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.

1. **Use case:** spatial proteomics.
2. Novel **computational biology research and developments** to acquire reliable biological knowledge.
3. **Behind the scenes:** software/data structures and open research practice.

Use case: spatial proteomics.

Cell organisation - regulation of protein localisation



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.

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

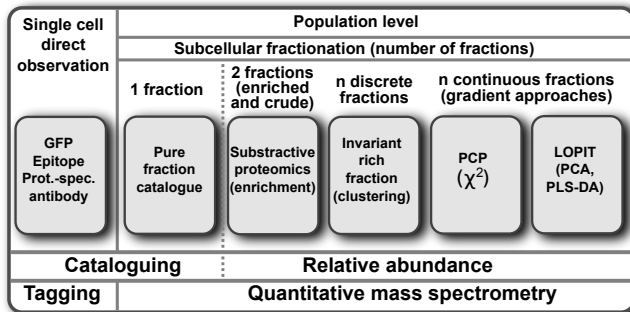


Figure: Organelle proteomics approaches ([Gatto et al., 2010](#))

Fusion proteins and immunofluorescence

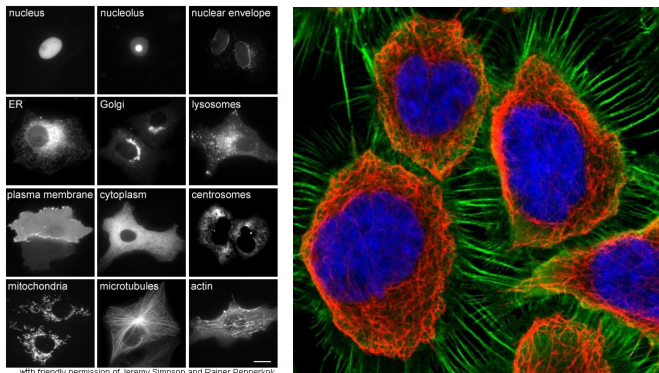


Figure: Targeted protein localisation. Example of discrepancies between IF and FPs as well as between FP tagging at the N and C termini (Stadler et al., 2013).

Spatial proteomics - How, experimentally

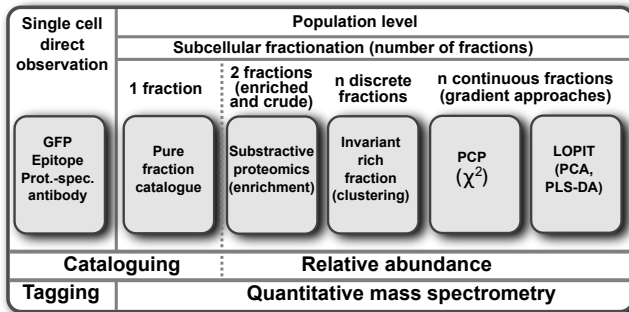
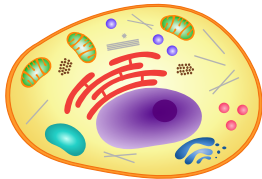


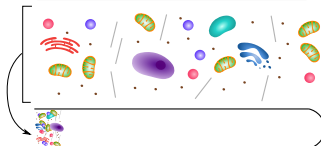
Figure: Organelle proteomics approaches ([Gatto et al., 2010](#)).

Gradient approaches: [Dunkley et al. \(2006\)](#), [Foster et al. \(2006\)](#).

Explorative/discovery approaches, [steady-state global localisation maps](#).

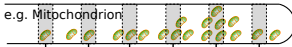


Cell lysis



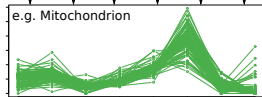
Fractionation/centrifugation

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}
⋮	⋮	⋮	⋮	⋮
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}	<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

Data analysis

Data analysis

- ▶ Visualisation (cluster, unsupervised learning)
- ▶ Classification (supervised learning)
- ▶ Novelty detection (semi-supervised learning)
- ▶ Data integration (transfer learning)
- ▶ Multi-localisation (Bayesian spatial proteomics)
- ▶ Spatial dynamics

To uncover and understand biology

Visualisation

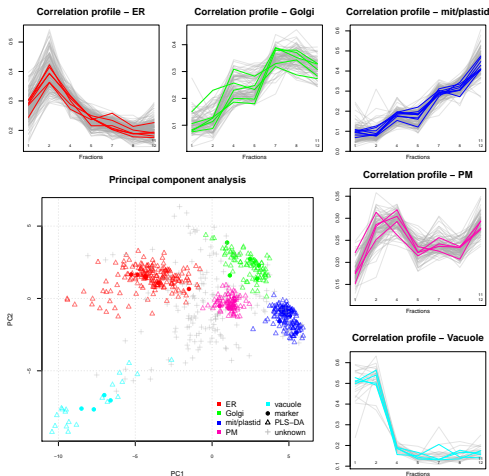


Figure: From Gatto et al. (2010), *Arabidopsis thaliana* data from Dunkley et al. (2006)

Supervised Machine Learning

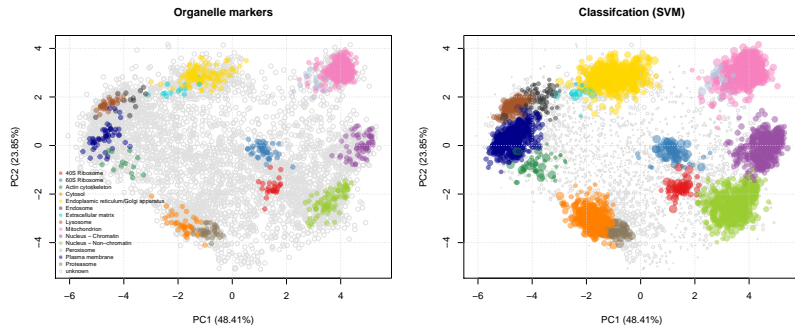
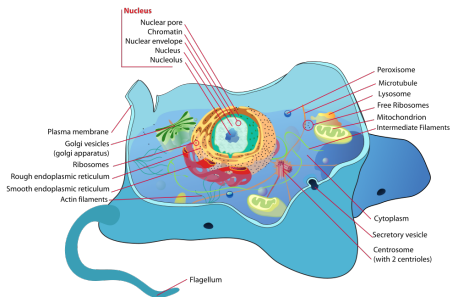
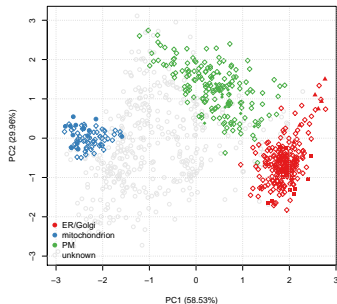


Figure: Support vector machines classifier (after 5% FDR classification cutoff) on the embryonic stem cell data from [Christoforou et al. \(2016\)](#).

Novel **computational biology research and developments** to acquire reliable biological knowledge.

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

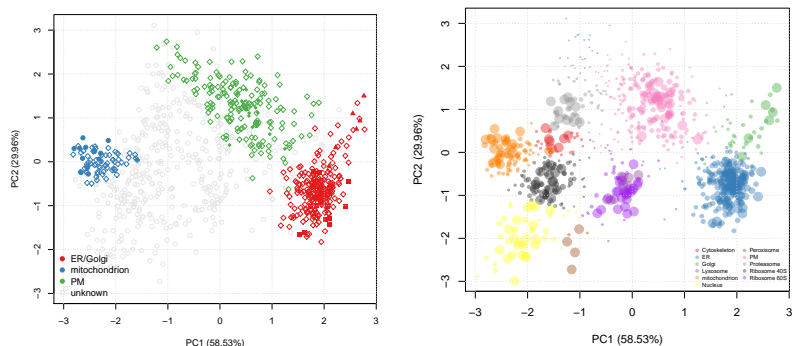


Figure: Left: Original *Drosophila* data from Tan et al. (2009). Right: After semi-supervised learning and classification, Breckels et al. (2013).

Computational advances: Transfer learning

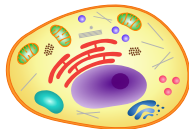
What about using **addition data**, such as annotations from the Gene Ontology (GO), sequence features (pseudo aminoacid composition), signal peptide, trans-membrane domains (length, number, ...), images (IF, FP), interaction data, prediction software, ...

- ▶ 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**

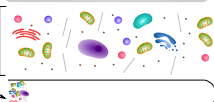
Transfer learning

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

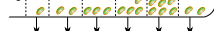


Cell lysis

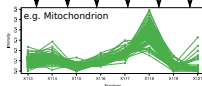


Fractionation/centrifugation

e.g. Mitochondrion

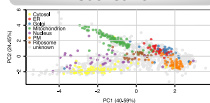


Quantitation/identification by mass spectrometry



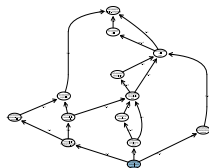
	X110	X114	X115	X116	X117	X118	X119	X121
CDMT7	0.180	0.150	0.090	0.167	0.277	0.1478	0.0380	0.0010
PR148	0.1014	0.200	0.000	0.180	0.207	0.000	0.0180	0.0777
CDMTA	0.107	0.201	0.000	0.180	0.200	0.1483	0.0200	0.0000
CDMTA	0.107	0.201	0.000	0.180	0.200	0.1483	0.0200	0.0000

Visualisation



Database query

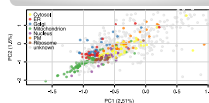
Extract GO CC terms



Convert terms to binary

	GO:0005822	GO:0005789	GO:0005783	GO:0005783
CDMT7	1	1	1	1
PR148	1	1	1	1
CDMTA	1	1	1	1
CDMTA	1	1	1	1

Visualisation



AUXILIARY DRY DATA

Transfer learning results

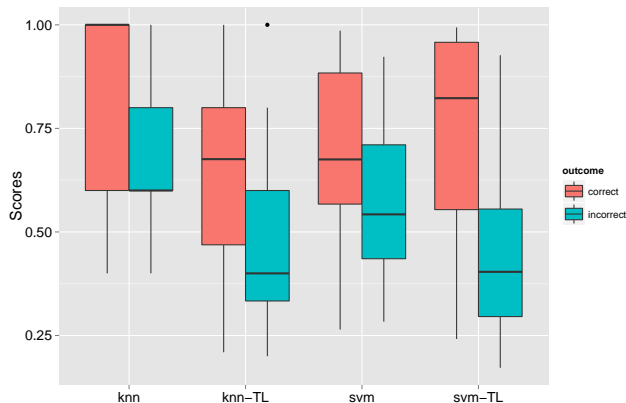
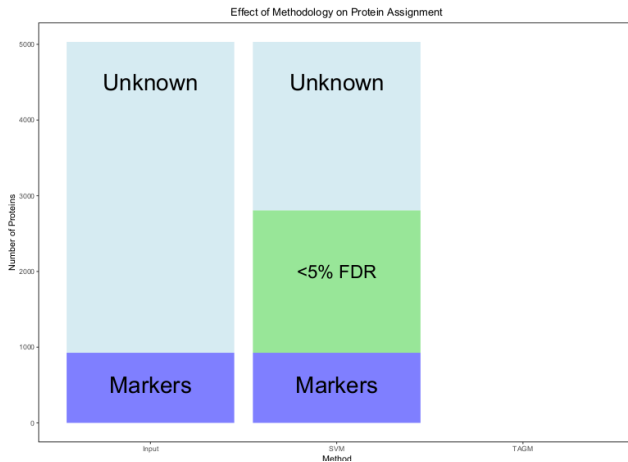


Figure: From Breckels et al. (2016) *Learning from heterogeneous data sources: an application in spatial proteomics*.

How much do we learn? How much do we miss?



A Bayesian Mixture Modelling Approach For Spatial Proteomics

- ▶ *T Augmented Gaussian Mixture model (TAGM)* is a **multivariate Gaussian generative model** for MS-based spatial proteomics data. It posits that each annotated sub-cellular niche can be modelled by a multivariate Gaussian distribution.

A Bayesian Mixture Modelling Approach For Spatial Proteomics

- ▶ *T Augmented Gaussian Mixture model (TAGM)* is a **multivariate Gaussian generative model** for MS-based spatial proteomics data. It posits that each annotated sub-cellular niche can be modelled by a multivariate Gaussian distribution.
- ▶ With the prior knowledge that many proteins are not captured by known sub-cellular niches, we augment our model with an **outlier component**. Outliers are often dispersed and thus this additional component is described by a heavy-tailed distribution: the multivariate Student's t-distribution, leading us to a *T Augmented Gaussian Mixture model*.

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.

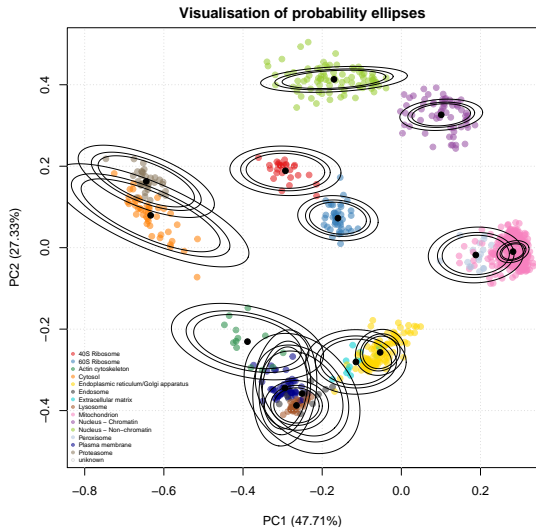


Figure: Illustration of how the TAGM model describes the pluripotent mouse embryonic stem cell data. Each ellipse contains a proportion of total probability of a particular multivariate Gaussian density. The outer ellipse contains 99% of the total probability whilst the middle and inner ellipses contain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

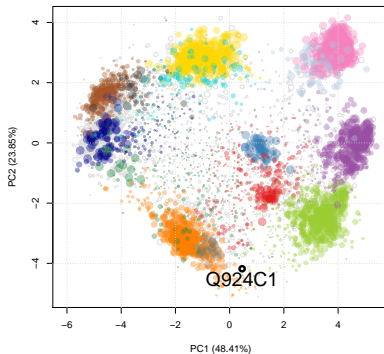
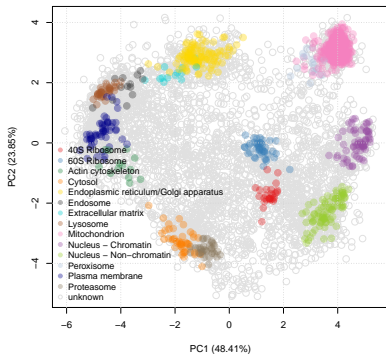
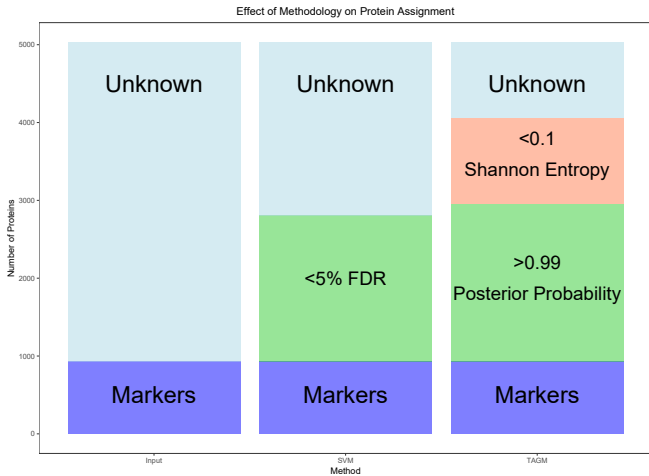


Figure: Assignment of proteins of *unknown* location to one of the annotated classes. The dots are scaled according to the protein assignment probabilities.



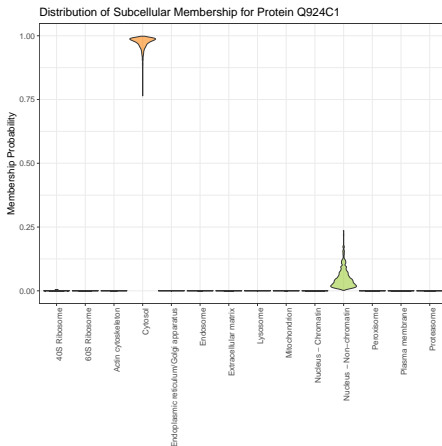
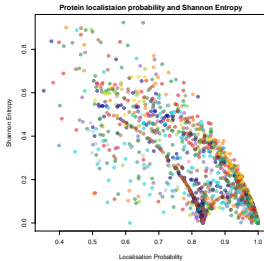
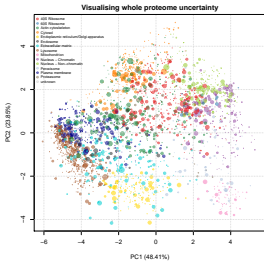
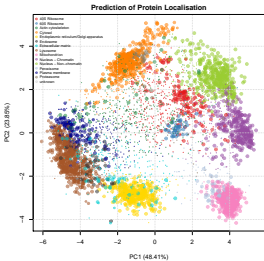


Figure: Exportin 5 (Q924C1) forms part of the micro-RNA export machinery, transporting miRNA from the nucleus to the cytoplasm for further processing. It then translocates back through the nuclear pore complex to return to the nucleus to mediate further transport between nucleus and cytoplasm. The model correctly infers that it most likely localises to the cytosol but there is some uncertainty with this assignment. This uncertainty is reflected in possible assignment of Exportin 5 to the nucleus non-chromatin and reflects the multi-location of the protein.

Whole sub-cellular proteome uncertainty



Spatial dynamics

Trans-localisation event during monocyte to macrophage differentiation

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

Data

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

Work lead by **Dr Claire Mulvey** at the Cambridge Centre for Proteomics.

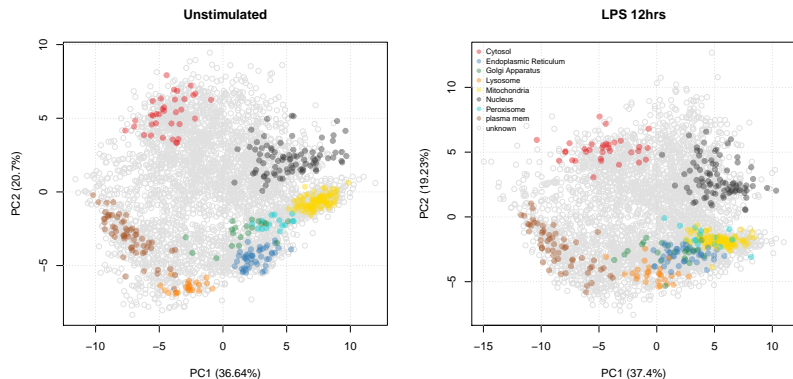


Figure: Spatial maps of unstimulated and LPS-treated cells (combined triplicates).

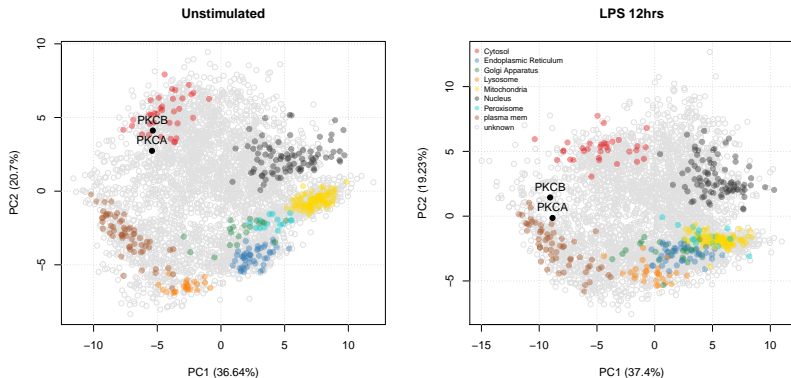


Figure: Relocation of Protein Kinase C α and β from the cytosol to the plasma membrane, **driving maturation into a differentiated macrophage phenotype.**

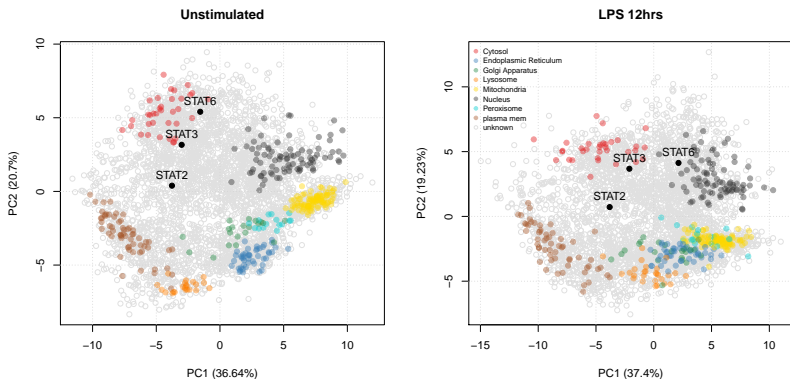


Figure: Relocation of Signal transducer and activator of transcription 6 (STAT6) from the cytosol to the Nucleus, **activating anti-bacterial and anti-viral-like response**. Validated by microscopy and see also [Chen et al. \(2011\)](#).

Behind the scenes: software/data structures and open research practice.

Beyond the figures¹

- ▶ 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/>

³between and within domains/software

Beyond the figures¹

- ▶ 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/>

³between and within domains/software

Open research: open source software

The screenshot shows the GitHub repository for **lgatto / pRoloc**. The repository is described as "A unifying bioinformatics framework for organelle proteomics" with a link to <http://lgatto.github.io/pRoloc/>. It has 2,051 commits, 10 branches, 25 releases, 1 environment, and 14 contributors. The repository includes a table of files and their last commit dates:

File	Last commit
R	fix to make work with devel 6 day
data	Merge branch 'master' into devel 2 year
inst	updated documentation 6 month
man	new logPosterior accessor 2 month
src	more C exporting fusa 10 month
tests	add test back 6 day
vignettes	automatic fix of indentation 6 day
_Rbuildignore	ignore .editorconfig when building 3 month
.editorconfig	fix tab with spaces and add editorconfig 4 month
gitignore	ignore docs, update news 5 month
travis.yml	Fix indentation 5 month
CONDUCT.md	Merge branch 'master' into devel 3 year
DESCRIPTION	bump devel version on gh 6 day
NAMESPACE	fix notes and warnings 7 day
NEWS	bump devel version on gh 6 day
NEWS.md	bump devel version on gh 6 day

The screenshot shows the Bioconductor website for the **pRoloc** package. The Bioconductor logo is at the top, with the tagline "OPEN SOURCE SOFTWARE FOR BIOINFORMATICS". The navigation bar includes "Home", "Install", and "Help". The main content area shows the package name **pRoloc** and the version **3.8**. It includes a table of statistics:

platforms	all	fork	254 / 164	posts	1 / 2 / 2 / 0	in BiOC	6 years
build	warnings	updated	< 1 week				

The DOI is [10.18129/B3.bioc.pRoloc](https://doi.org/10.18129/B3.bioc.pRoloc). The description states: "A unifying bioinformatics framework for spatial proteomics". The author is Laurent Gatto, Oliver Crook and Lisa M. Breckels. The citation is: Gatto L, Breckels LM, Wiecek S, Burger T, Lilley KS (2014). "Mass-spectrometry based spatial proteomics data analysis using pRoloc and pRolocdata." *Bioinformatics*.

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

Open and reproducible research

The figure displays three screenshots illustrating the open and reproducible research workflow for the Gatto et al. (2018) study.

Left Screenshot (GitHub): Shows the repository `lgatto/QSep-manuscript`. The file tree includes:

- `data`: add marker transfer code/figs
- `figure`: Update for bioRxiv
- `pipeline`: add cover letter 2
- `travis.yml`: add travis file
- `Makefile`: addressing more reviewers comments
- `README.md`: Update README.md
- `cover.pdf`: add cover letter
- `cover.tex`: add cover letter
- `cover2.pdf`: add cover letter 2
- `chlam_rdx`: qsep assessment section with rib cluster sims
- `hdm_rdx`: qsep assessment section with rib cluster sims
- `mk.R`: Calculate qsep distribution medians
- `reknach_qsep.pdf`: incorporate Kathryn and Lisa's comments
- `reknach_qsep.pdf`: incorporate Kathryn and Lisa's comments
- `reknach_qsep.pdf`: minor updates and change marker transfer paragraph
- `qsep.R`: fix table
- `qsep.Rnw`: Update for bioRxiv
- `qsep.bib`: changes to new part in col
- `qsep.pdf`: Update for bioRxiv
- `qsep.tex`: Update for bioRxiv
- `sim.pdf`: incorporate Kathryn and Lisa's comments
- `sim.R`: incorporate Kathryn and Lisa's comments

Middle Screenshot (bioRxiv): Shows the preprint page for "Assessing sub-cellular resolution in spatial proteomics experiments" by Laurent Gatto^{1,2,3,4}, Lisa M. Breckels^{1,2}, and Kathryn S. Lilley². The page includes a "New Results" badge, a "View current version of this article" link, and a "Comment on this paper" button.

Right Screenshot (Current Opinion in Chemical Biology): Shows the published paper in *Current Opinion in Chemical Biology*, Volume 48, February 2019, Pages 123-149. The page includes an "Abstract" section and a "Show more" link.

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

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

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

Conclusions

- ▶ Protein sub-cellular localisation: technologies (hyperLOPIT) and opportunities.
- ▶ Reliance on computational biology and dedicated software (pRoLoc *et al.*) to interpret data and acquire biological knowledge.
- ▶ Rigorous computational infrastructure and sound data analysis and interpretation is a **long term investment**.

References |

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

References 11

- 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.
- 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**, (U of Cambridge): spatial proteomics, machine learning, software.
- ▶ **Dr Sebastian Gibb** and **Dr Johannes Rainer**: MS and proteomics software.
- ▶ Prof Kathryn Lilley (U of Cambridge), Dr Claire Mulvey, (CRUK Cambridge Institute): data.
- ▶ Funding: BBSRC, Wellcome Trust

Thank you for your attention