

Probabilistic modelling of protein sub-cellular localisation

Laurent Gatto

`laurent.gatto@uclouvain.be`

`http://lgatto.github.io/about`

de Duve Institute – UCLouvain

14 May 2019 – Compomics

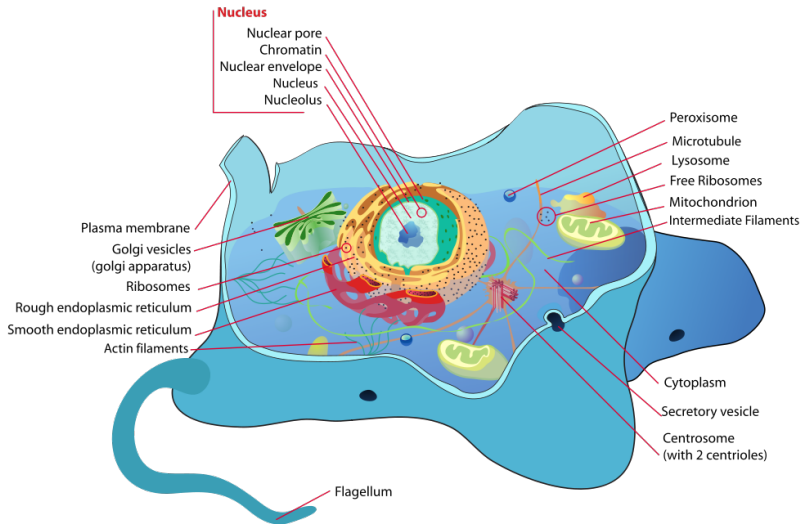
Abstract

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:** Localisation and sequestration of proteins within sub-cellular niches is a fundamental mechanism for the post-translational regulation of protein function.
- ▶ **Re-localisation:** *differentiation* stem cells, *activation* of biological processes.
- ▶ **Mis-localisation:** Disruption of the targeting/trafficking process alters proper sub-cellular localisation, which in turn perturb the cellular functions of the proteins.

Spatial proteomics - How, experimentally

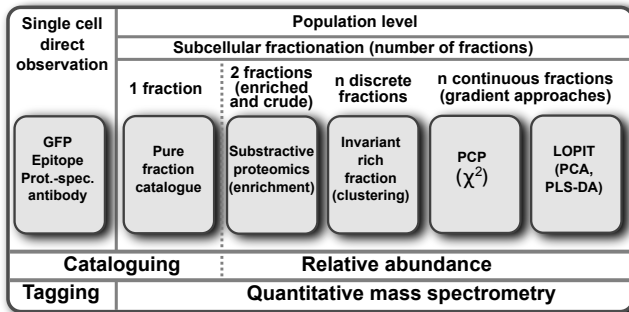
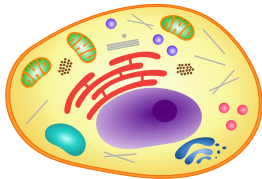


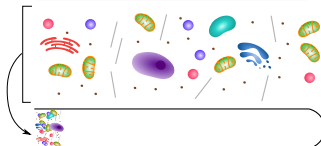
Figure: Organelle proteomics approaches ([Gatto et al., 2010](#)).

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

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

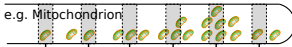


Cell lysis



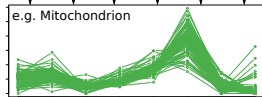
Fractionation/centrifugation

e.g. Mitochondrion



Quantitation/identification
by mass spectrometry

e.g. Mitochondrion



Quantitation data

	Fraction ₁	Fraction ₂	...	Fraction _L
x₁	x _{1,1}	x _{1,2}	...	x _{1,L}
x₂	x _{2,1}	x _{2,2}	...	x _{2,L}
x₃	x _{3,1}	x _{3,2}	...	x _{3,L}
⋮	⋮	⋮	⋮	⋮
x_i	x _{i,1}	x _{i,2}	...	x _{i,L}
⋮	⋮	⋮	⋮	⋮
x_N	x _{N,1}	x _{N,2}	...	x _{N, L}

Quantitation data and organelle markers

	Fraction ₁	Fraction ₂	...	Fraction _L	markers
x_1	$x_{1,1}$	$x_{1,2}$...	$x_{1,L}$	unknown
x_2	$x_{2,1}$	$x_{2,2}$...	$x_{2,L}$	<i>loc₁</i>
x_3	$x_{3,1}$	$x_{3,2}$...	$x_{3,L}$	unknown
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
x_i	$x_{i,1}$	$x_{i,2}$...	$x_{i,L}$	<i>loc_k</i>
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
x_N	$x_{N,1}$	$x_{N,2}$...	$x_{N,K}$	unknown

Data analysis

Visualisation

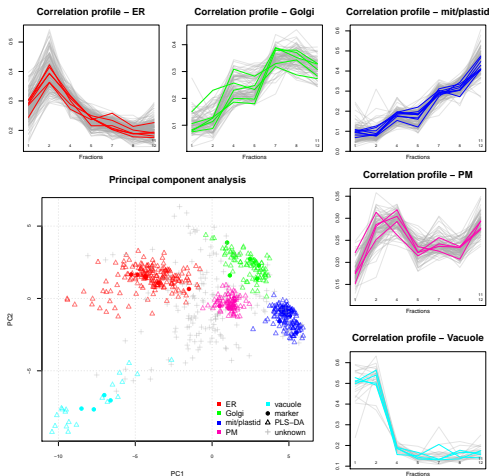


Figure: From Gatto et al. (2010), *Arabidopsis thaliana* data from Dunkley et al. (2006)

Supervised Machine Learning

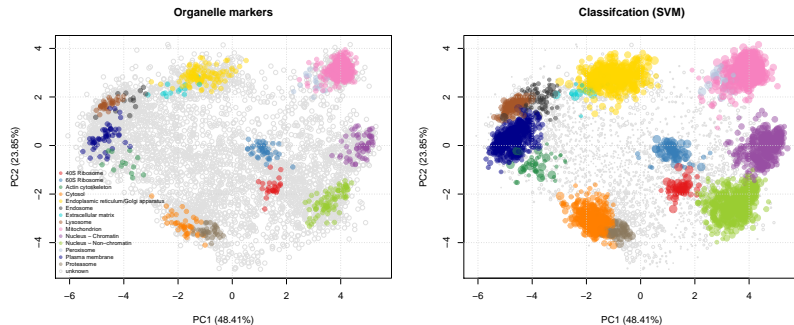
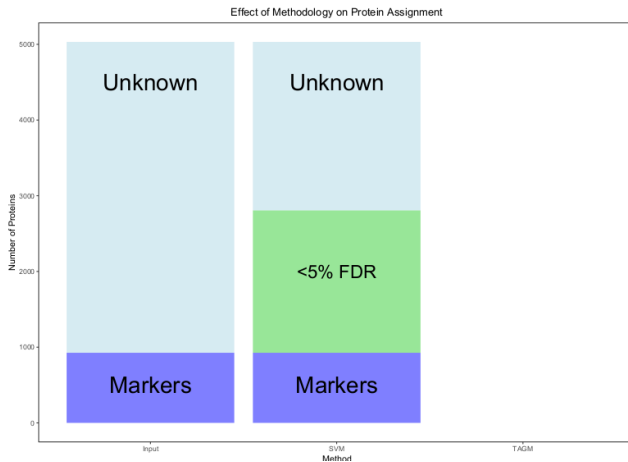


Figure: Support vector machines classifier (after 5% FDR classification cutoff) on the embryonic stem cell data from [Christoforou et al. \(2016\)](#).

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



RESEARCH ARTICLE

A Bayesian mixture modelling approach for spatial proteomics

Oliver M. Crook^{1,2,3}, Claire M. Mulvey², Paul D. W. Kirk³, Kathryn S. Lilley², Laurent Gatto^{1,2*}

1 Computational Proteomics Unit, Department of Biochemistry, University of Cambridge, Cambridge, UK, **2** Cambridge Centre for Proteomics, Department of Biochemistry, University of Cambridge, Cambridge, UK, **3** MRC Biostatistics Unit, Cambridge Institute for Public Health, Cambridge, UK

✉ Current address: de Duve Institute, UCLouvain, Brussels, Belgium

* laurent.gatto@uclouvain.be



OPEN ACCESS

Citation: Crook OM, Mulvey CM, Kirk PDW, Lilley KS, Gatto L (2018) A Bayesian mixture modelling approach for spatial proteomics. PLoS Comput Biol 14(11): e1006516. <https://doi.org/10.1371/journal.pcbi.1006516>

Editor: Christine Vogel, NYU, UNITED STATES

Received: May 23, 2018

Accepted: September 17, 2018

Published: November 27, 2018

Copyright: © 2018 Crook et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Analysis of the spatial sub-cellular distribution of proteins is of vital importance to fully understand context specific protein function. Some proteins can be found with a single location within a cell, but up to half of proteins may reside in multiple locations, can dynamically re-localise, or reside within an unknown functional compartment. These considerations lead to uncertainty in associating a protein to a single location. Currently, mass spectrometry (MS) based spatial proteomics relies on supervised machine learning algorithms to assign proteins to sub-cellular locations based on common gradient profiles. However, such methods fail to quantify uncertainty associated with sub-cellular class assignment. Here we reformulate the framework on which we perform statistical analysis. 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, with Bayesian computation performed using the expectation-maximisation (EM) algorithm, as well as Markov-chain Monte-Carlo (MCMC). Our methodology allows proteome-wide uncertainty quantification, thus adding a further layer to the analysis of spatial proteomics. Our framework is flexible, allowing many different systems to be analysed and reveals new modelling opportunities for spatial proteomics. We find our methods perform competitively with current state-of-the-art machine learning methods, whilst simultaneously providing

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**, thus adding a further layer to the analysis of spatial proteomics.

We initially model the distribution of profiles associated with proteins that localise to the k -th component as multivariate normal with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$, so that:

$$\mathbf{x}_i | z_i = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (1)$$

We initially model the distribution of profiles associated with proteins that localise to the k -th component as multivariate normal with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$, so that:

$$\mathbf{x}_i | z_i = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (1)$$

We extend it by introducing an additional *outlier component*. To do this, we augment our model by introducing a further indicator latent variable ϕ . Each protein \mathbf{x}_i is now described by an additional variable ϕ_i , with $\phi_i = 1$ indicating that protein \mathbf{x}_i belongs to a organelle derived component and $\phi_i = 0$ indicating that protein \mathbf{x}_i is not well described by these known components. This outlier component is modelled as a multivariate T distribution with degrees of freedom κ , mean vector \mathbf{M} , and scale matrix V .

$$\mathbf{x}_i | z_i = k, \phi_i \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{\phi_i} \mathcal{T}(\kappa, \mathbf{M}, V)^{1-\phi_i} \quad (2)$$

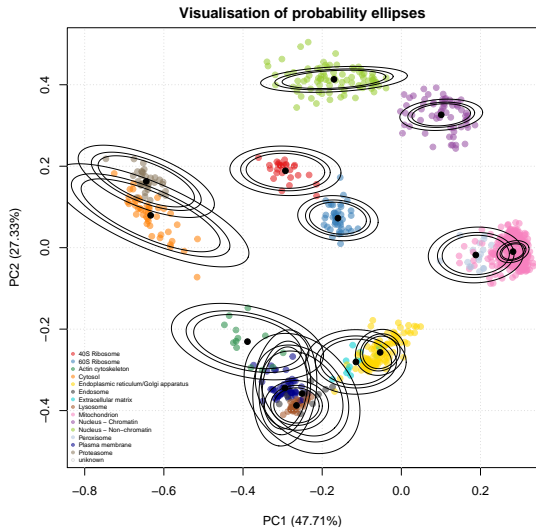


Figure: Illustration of how the TAGM model describes the pluripotent mouse embryonic stem cell data. Each ellipse contains a proportion of total probability of a particular multivariate Gaussian density. The outer ellipse contains 99% of the total probability whilst the middle and inner ellipses contain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

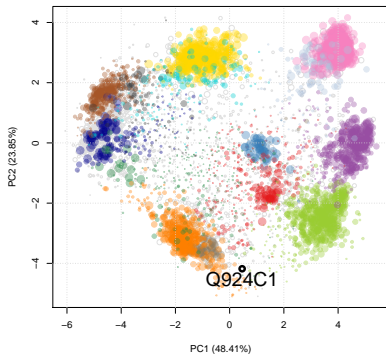
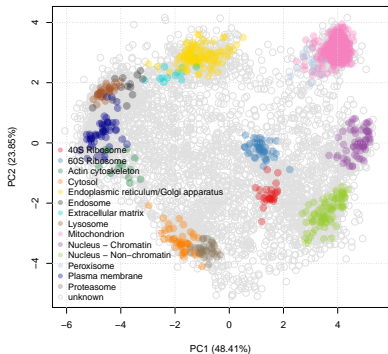
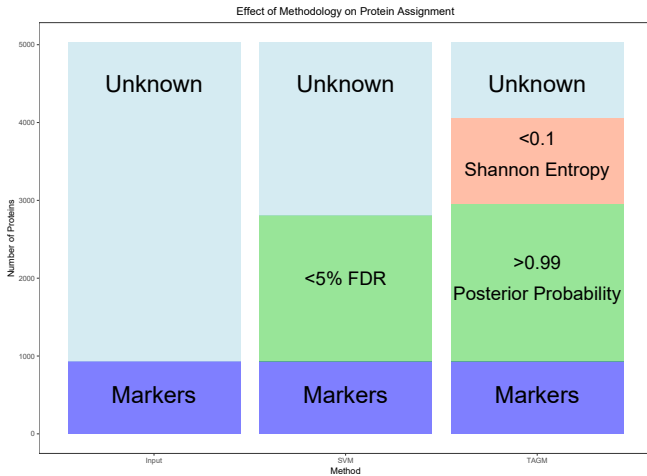


Figure: Assignment of proteins of *unknown* location to one of the annotated classes. The dots are scaled according to the protein assignment probabilities.



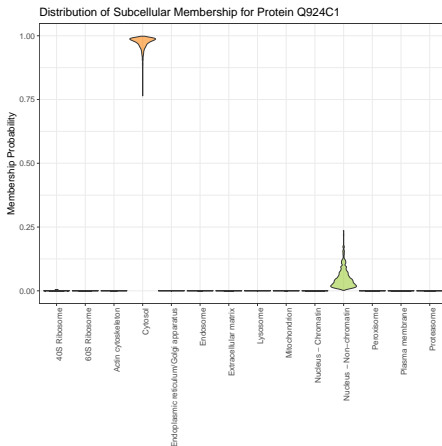
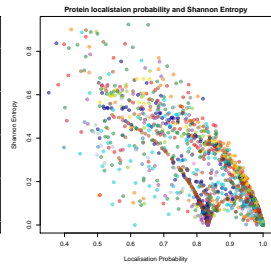
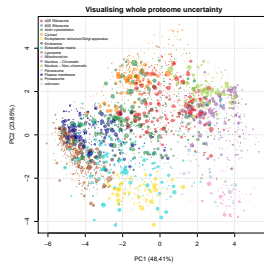
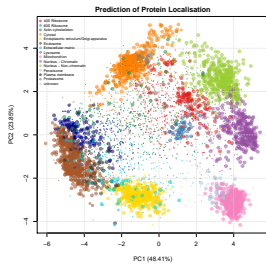


Figure: Exportin 5 (Q924C1) forms part of the micro-RNA export machinery, transporting miRNA from the nucleus to the cytoplasm for further processing. It then translocates back through the nuclear pore complex to return to the nucleus to mediate further transport between nucleus and cytoplasm. The model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



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 description is "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 file tree with folders like `R`, `data`, `inst`, `man`, `src`, `tests`, `vignettes`, and files like `_Rbuildignore`, `.editorconfig`, `gitignore`, `travis.yml`, `CONDUCT.md`, `DESCRIPTION`, `NAMESPACE`, `NEWS`, and `NEWS.md`. Each file has a commit message and a timestamp.

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 statistics: 254 forks, 164 stars, 17 posts, 2/2/0 in BioC, and 6 years. The DOI is [10.18129/B9.bioc.pRoloc](https://doi.org/10.18129/B9.bioc.pRoloc). The description states: "The pRoloc package implements machine learning and visualisation methods for the analysis and interrogation of quantitative mass spectrometry data to reliably infer protein sub-cellular localisation." The authors are 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

Iqbal / QGep-manuscript

Code Issues Pull requests Insights Settings

Assessing sub-cellular resolution in spatial proteomics experiments

proteomics spatial proteomics protein proteobacteria Manage topics

95 commits p1 branch 8 releases

Search master New pull request Create new file Upload files

File Name	Description
igbio Update README.md	
fig_data	add marker transfer code/figs
fig_figure	Update for bioRxiv
q_globe	add cover letter 2
trans.pyd	add trans file
Makelife	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
e14m_rdx	qsep assessment section with rib-cluster sims
hdm.rdx	qsep assessment section with rib-cluster sims
mark.R	Calculate qsep distribution means
markswitch-pca.pdf	Incorporate Kallrityn and Lucks comments
markswitch-qsep.pdf	Incorporate Kallrityn and Lucks comments
markswitch.rdx	minor updates and change marker transfer paragraph
fix_table	fix table
qsep.R	Update for bioRxiv
qsep/Rnew	changes to new part in ccl
qsep.kids	Update for bioRxiv
qsep.pdf	Update for bioRxiv
qsep.tex	Update for bioRxiv
sims.pdf	Incorporate Kallrityn and Lucks comments
sims.R	Incorporate Kallrityn and Lucks comments



Cold Spring
Harbor
Laboratory



THE PREPRINT SERVER FOR BIOLOGY

[HOME](#) | [ABOUT](#) | [SUBMIT](#) | [ALERTS/RSS](#) | [CHANNELS](#)

Advanced Search

New Results

View current version of this article

Assessing sub-cellular resolution in spatial proteomics experiments

Currently on this page

 Laurent Gatto, Lisa M Bruckers, Kathryn S Lilley

doi: <https://doi.org/10.1101/317780>

New published in *Current Opinion in Chemical Biology*; doi: [10.1016/j.cop.2018.11.015](https://doi.org/10.1016/j.cop.2018.11.015)

Abstract

Info-History

Metrics

Preview PDF

Abstract


The sub-cellular localisation of a protein is vital in defining its function, and a protein's mis-localisation is known to lead to adverse effect. As a result, numerous experimental techniques and datasets have been published, with the aim of deciphering the localisation of proteins at various scales and resolutions, including high profile mass spectrometry-based efforts. Here, we present a meta-analysis assessing and comparing the sub-cellular resolution of 29 such mass spectrometry-based spatial proteomics experiments using a newly developed tool termed CSep. Our goal is to provide a simple quantitative review of how well spatial proteomics resolve the sub-cellular niches they describe to inform and guide developers and users of such methods.

Copyright

The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Twitter

Twitter




ScienceDirect

Outline

Download

Share


Export



ELSEVIER

Current Opinion in Chemical Biology

Volume 48, February 2019, Pages 123–149



Get rights and content

open access

Assessing sub-cellular resolution in spatial proteomics experiments

Laurent Gatto^{1,*,1,49}, Lisa M. Breckels^{2,2}, Kathryn S. Lilley²

<https://doi.org/10.1016/j.copbio.2018.11.015>

Under a Creative Commons license

Abstract

The **sub-cellular localisation** of a protein is vital in defining its function, and a protein's mis-localisation is known to lead to adverse effect. As a result, numerous **experimental techniques** and datasets have been published, with the aim of deciphering the localisation of proteins at various scales and resolutions, including high profile mass spectrometry-based efforts. Here, we present a meta-analysis assessing and comparing the sub-cellular resolution of 29 such mass spectrometry-based spatial **proteomics** experiments using a newly developed tool termed **QSeq**. Our goal is to provide a simple quantitative report of how well temporal proteomics resolve the sub-cellular niches they describe and to inform and guide developers and users of such methods.

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, statistics 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 |

- 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.
- 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 Wicczorek, T Burger, and K S Lilley. Mass-spectrometry based spatial proteomics data analysis using pRoloc and pRolocdata. *Bioinformatics*, Jan 2014a.
- L Gatto, LM Breckels, T Burger, DJ Nightingale, AJ Groen, C Campbell, N Nikolovski, CM Mulvey, A Christoforou, M Ferro, and KS Lilley. A foundation for reliable spatial proteomics data analysis. *MCP*, 13(8):1937–52, Aug 2014b.
- Laurent Gatto, Lisa M Breckels, and Kathryn S Lilley. Assessing sub-cellular resolution in spatial proteomics experiments. *bioRxiv*, 2018. doi: 10.1101/377630.
- W Huber, V J Carey, R Gentleman, S Anders, M Carlson, B S Carvalho, H C Bravo, S Davis, L Gatto, T Girke, R Gottardo, F Hahne, K D Hansen, R A Irizarry, M Lawrence, M I Love, J MacDonald, V Obenchain, A K Oleś, H Pagès, A Reyes, P Shannon, G K Smyth, D Tenenbaum, L Waldron, and M Morgan. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods*, 12(2):115–21, Jan 2015. doi: 10.1038/nmeth.3252.

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