Assessing sub-cellular resolution in spatial proteomics experiments

Laurent Gatto* and Lisa M. Breckels

Computational Proteomics Unit, University of Cambridge, UK

August 22, 2016

Abstract

A meta-analysis comparing and assessing the sub-cellular resolution of spatial proteomics experiments.

1 Introduction

In biology, the localisation of a protein to its anticipated sub-cellular niche is an necessary condition for it to assume its biological function. Indeed, the localisation of a protein will confer a specific biochemical environment and a unique set of interaction partners. As a result, the same protein can assume different functions in different biological contexts and its mis-localisation can lead to adverse effects.

Spatial proteomics is the systematic and high-throughput study of protein sub-cellular localisation. A wide range of techniques [Gatto et al., 2010] and computational methods [Gatto et al., 2014a] to confidently infer the localisation of thousands of proteomics. Most techniques rely on some form of sub-cellular fractions using differential centrifugation or separation along density gradients and the quantitative assessment of relative protein occupancy profiles in these sub-cellular fractions. Reciprocally, a wide ranging of

*lg390@cam.ac.uk

computational methods have been applied, ranging from clustering [**?**], classification (reviewed in [Gatto et al., 2014a], semi-supervised learning [Breckels et al., 2013] and, more recently, transfer learning [Breckels et al., 2016].

Despite these advances, there is surprisingly little agreement in the community as to what constitutes are reliable spatial proteomics experiment, i.e a dataset that generates confident protein assignment results. Is is however implicit that reliability and trust in the results is dependent on adequate sub-cellular resolution, i.e. *enough* separation between the different sub-cellular niches under study. And yet, every spatial proteomics publication will claim to have obtain from satisfactory to excellent resolution.

The importance of adequate sub-cellular resolution reaches beyong the generation of reliable static spatial maps. It is a necessary quality of the data to consider tackling more subtle sub-cellular patterns such as multi- and trans-localisation, i.e. the localisation of proteins in multiple sub-cellular niches and the relocation of proteins upon perturbation [Gatto et al., 2014a].

In this work, we describe how to understand and interprete widely used dimensionality reduction methods and visualisations of spatial proteomics data to critically assess their resolution and propose a simple, yet effective method to quantitatively measure resolution and compare it across different experiments. Our recommendations should be useful to spatial proteomics practitioners, to assess the sub-cellular resolution of their experiments and compare it to similar studies while setting up and optimising their experiments, as well biologists interested in critically assessing spatial proteomics studies and their claims.

2 Spatial proteomics datasets

For this meta-analysis, we make use of 12 spatial proteomics datasets (table 1). These data represent a diverse range of species, instruments and methodologies.

We have applied minimal post-processing to the data and have used, as far as possible, the data and annotation provided by the original authors. The data from Foster et al. [2006] has been annotated using the curated marker list from Christoforou et al. [2016], as only a limited number of markers was provided by the authors[1].

[1]This results from the fact that they used a simple distance measurement, termed χ^2 against few markers to base their assignments

Data	Proteins	Clusters	Title
E14TG2aS1	1109	10	LOPIT experiment on Mouse E14TG2a Embryonic Stem Cells from Breckels et al. (2016)
andy2011	1371	12	LOPIT experiment on Human Embryonic Kidney fibroblast cells from Breckels et al. (2013)
dunkley2006	689	9	LOPIT data from Dunkley et al. (2006)
foster2006	1555	8	PCP data from Foster et al. (2006)
groen2014cmb	424	7	LOPIT experiments on Arabidopsis thaliana roots, from Groen et al. (2014)
hall2009	1090	5	LOPIT data from Hall et al. (2009)
hyperLOPIT2015	5032	14	hyperLOPIT experiment on Mouse E14TG2a embryonic stem cells from Christoforou et al. (2016)
hyperLOPIT2015ms2	7114	14	hyperLOPIT experiment on Mouse E14TG2a embryonic stem cells from Christoforou et al. (2016)
itzhak2016stcSILAC	5265	12	Data from Itzhak et al. (2016)
nikolovski2012imp	1385	7	Meta-analysis from Nikolovski et al. (2012)
nikolovski2014	1385	8	LOPIMS data from Nikolovski et al. (2014)
tan2009r1	888	11	LOPIT data from Tan et al. (2009)

Table 1: Summary of the datasets used in this study. See also figure 5 for the individual PCA plots.

In addition, for dimensionality reduction and visualisation, we have systematically replaced missing values by zeros. When caluclation distances between protein profiles, however, missing values were retained.

It is important to highlight that not all experiments used in this study have as main goal the generation of a global sub-cellular map. While the works of Dunkley et al. [2006] (*Mapping the Arabidopsis organelle proteome*), Hall et al. [2009] (*Mapping organelle proteins and protein complexes in Drosophila melanogaster*) and more recently Christoforou et al. [2016] (*A draft map of the mouse pluripotent stem cell spatial proteome*) and Itzhak et al. [2016] (*Global, quantitative and dynamic mapping of protein subcellular localization*) explicitly state such a goal, other experiments such as Groen et al. [2014] (*Identification of trans-golgi network proteins in Arabidopsis thaliana root tissue*) or Nikolovski et al. [2014] (*Label free protein quantification for plant Golgi protein localisation and abundance*) have a much more targetted goal (trans-Golgi and Golgi apparatus, respectively). Hence, it is important to keep the overall aim of the studies in mind when assessing their resolution.

3 Assessment

Sub-cellular diversity

An important assessment that provides an important indication of the resolution of the data concerns the number and diversity of sub-cellular niches that are annotated. In the 12 datasets used in this study, this number ranged from 5 (dataset hall2009) to 14 (dataset hyperLOPIT2015). These numbers can be compared to about 25 different organelles documented in all datasets, which are still underestimating the biological sub-cellular diversity.

Dimensionality reduction and visualisation

Principal component analysis (PCA) is a widely used dimensionality rediction technique is spatial proteomics. It project the protein occupancy profiles into a new space in such a way as to maximise the spread of all points (i.e. labelled and unlabelled proteins) along the first new dimension (principal component, PC). The second PC is then chosen to be perpendicular to the first one while still maximising the overall variability. Each PC (there are as many as there are fractions) accounts for a percentage of the total variability

and it is not uncommon, in well exectuted experiments, that the two first PCs summarise over 70% of the total variance in the data.

By firstly summerising the complete data along PC1 and PC2 (and, possibly, another component of interest), it becomes possible to visualise the complete dataset in a single figure (as opposed to individual sets of profiles - see for example figures 5 in Gatto et al. [2010]). In a first instance, it is advised to visualise the data without annotation to confirm the presence of discrete clusters, i.e. dense clouds of points that are well separated from the rest of the data (see for example data from Christoforou et al. [2016] on figure 1, left). This can be highlighted by using transparency to reflect density (figure 1, centre) or binning plot regions (figure 1, right).

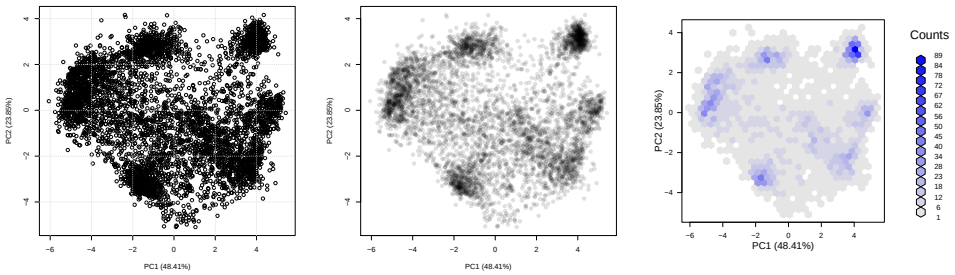

Figure 1: Unsupervised visualisation of spatial resolution

In figure 2 we compare three datasets to illustrate differnt levels of cluster density and separation. The figure on the left the the hyperLOPIT data from Christoforou et al. [2016] (as on figure 1) that used synchronous precursor selection (SPS) MS^3 on an Orbitrap Fusion. The middle figure represents the same experiment and same proteins, analysed using conventional MS^2, illustrating the effect of reduced quantitation accuracy. Finally, on the left, an experiment with considerable less resolution [Hall et al., 2009].

Considering that the aim of sub-cellular fractions is to maximise separation of (some and most) sub-cellular niches, one would expect these sub-cellular clusters to be separated optimally in a successfull spatial proteomics experiment. In PCA space, this would equate to the location of the annotated spatial clusters along the periphery of the data points. In other words, the maximum variability of a successfull spatial proteomics experiments should be reflected by the separation of the spatial clusters.

Another dimensionality rediction method that is worth mentioning here is

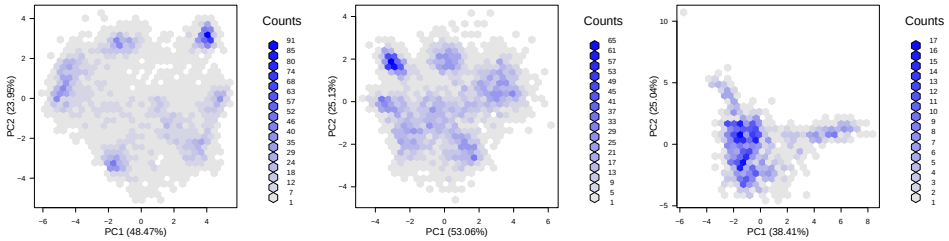

Figure 2: Comparing the cluster density and separation of a very good separation from best (left) to worst (right).

linear discriminant analysis. LDA will project the protein occupancy profiles in a new set of dimensions using as criteria the maximal separation of marker classes by maximising the between class variance to the within calss variance ratio. As opposed to the *unsupervised* PCA, the *supervised* LSA should not be used as an experiment quality control, but can be useful to assess if one or more organelles have been preferentially separated.

Quantifying resolution

While visualisation of spatial proteomics data remains essential to assess the resolution, and hence the success, of a spatial proteomics experiment, it is useful to be able to objectively quantify the resolution and directly compare different experiments. Here, we present a new infrastructure to quantify the resolution in spatial proteomics experiments, that relies on the comparison of the average euclidean distance within and between sub-cellular clusters.

As illustrated on the heatmaps in figure 3, these distances always refer to one reference marker cluster. The raw distance matrix (figure 3, top-left) is symmetrical (i.e. the distance beetween cluster 1 and 2 is the same as between cluster 2 and 1). Within distances are generally the smallest ones, except when two clusters overlap, as the lysosome and endosome in our example. To enables the comparion of these distances within and between experiements (see section 4 for the lattes), we further divide each distance be the reference within cluster average distance (figure 3, top-right). This thus informs is as how much the average distance between cluster 1 and 2 is greater than the average distance within orgnalle 1 (i.e. the tightness of that

cluster). At this stage, the distance matrix is not symmetrical anymore. To faciliate the comparison of distances between organelles, we also illustrate the distance distributions as boxplots (figure 3, bottom).

The rational behind these measures is as follows. Intuitively, we assess resolution by using speration of clusters (formalised by the average distance between two clusters) and tighness of one cluster (formalised by the average within cluster distance). Ideal sub-cellular fractionation would yield tight and distant clusters, represented by a large normalised between cluster distances on figure 3.

4 Comparative study

We now apply the quantitative assessment of spatial resolution described in section 3 to compare the 12 experiments presented in section 2. On figure 4, we display the global average normalised between cluster distances for all spatial clusters on a single boxplot per datasets. The datasets have been ordered using the experiment-wide median between distance. It is important to refer back to always refer the original data when considering summerising metrics like these, to put the resolution into context; the annotated and density PCA plots discussed in section 3 are also provided in figures 5 and 6.

5 Conclusions

It is important to highlight the importance and effect of marker definition on estimating and assessing the resolution of spatial proteomics experiments and, of course, assignment of proteins to their most likely sub-cellular compartments. In this work, we have used the markers provided by the original authors (except for Foster et al. [2006] to assess the data as originally presented.

TODO: Number and tighness of clusters. Cluster boundaries.

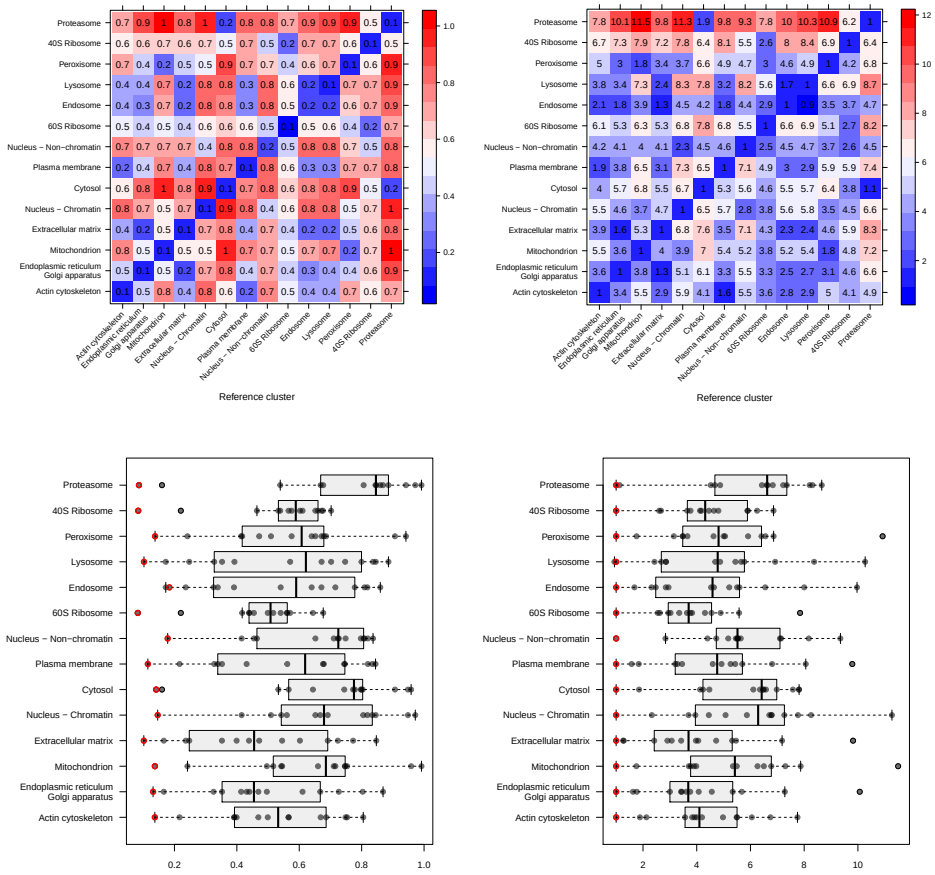

Figure 3: Quantifying resolution of Christoforou et al. [2016]. The heatmaps at the top illustrate the raw (left) and average normalised (right) within (along the diagonal) and between euclidean cluster distances. The boxplots at the bottom summarise these same values (raw on the left, normalised on the right) to enable easier comparison between clusters, where the within distances are highligted in red.

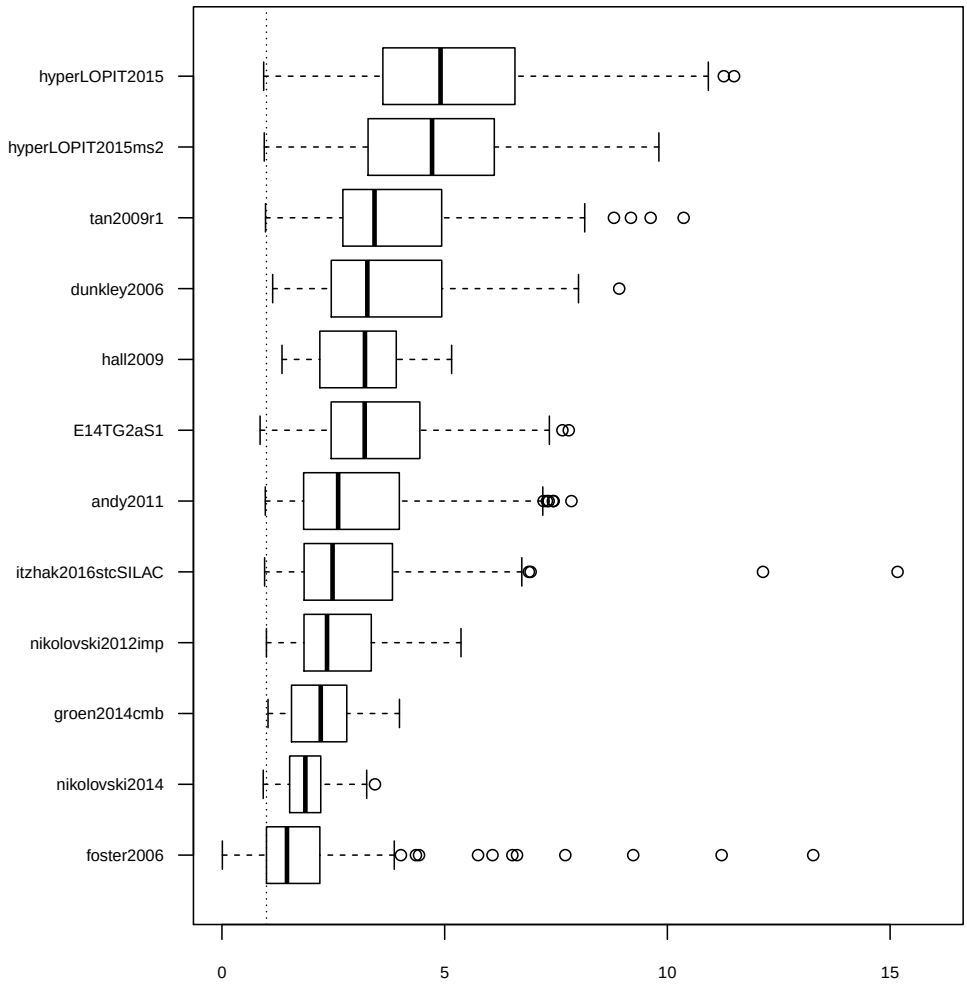

Figure 4: Quantitative separation assessment using experiment-wide normalised between cluster distances.

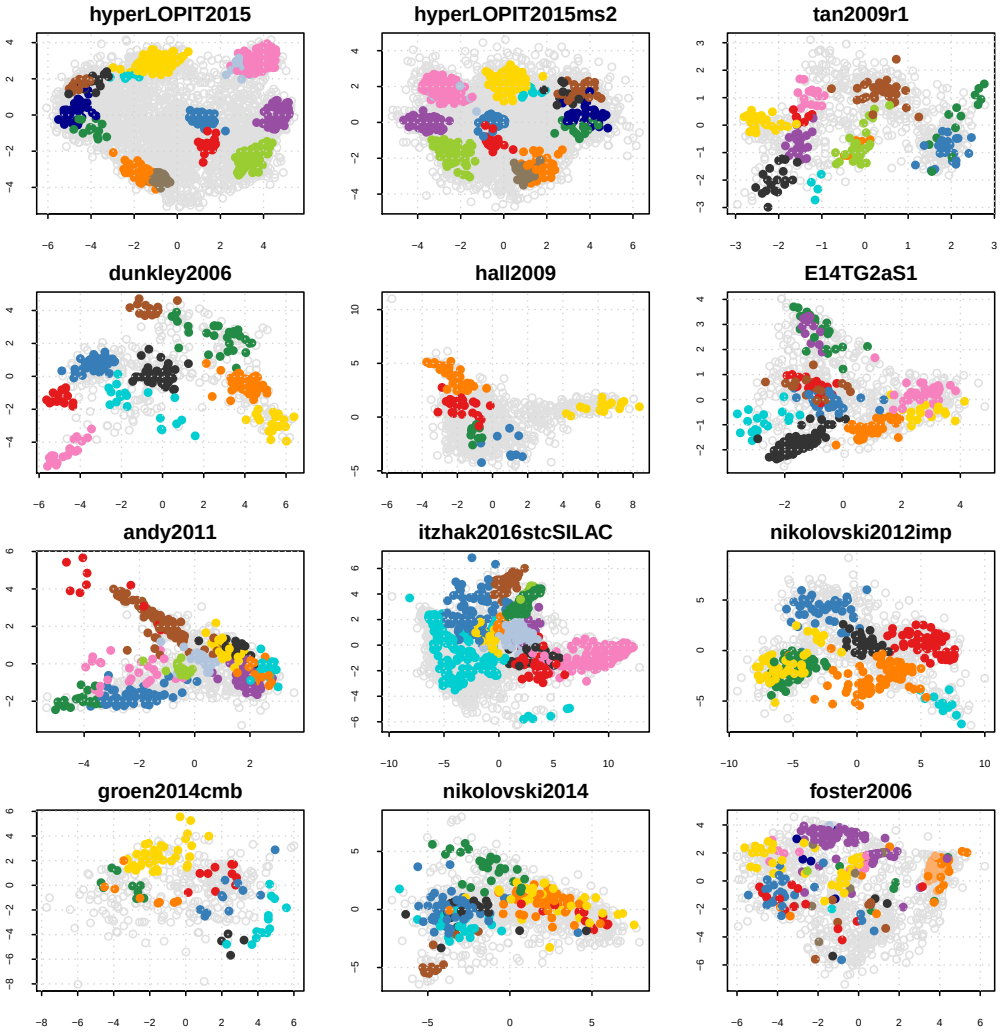

Figure 5: PCA plots for the 12 experiments used in this study. The experiments are ordered according to the median average between cluster distance (see figure 4).

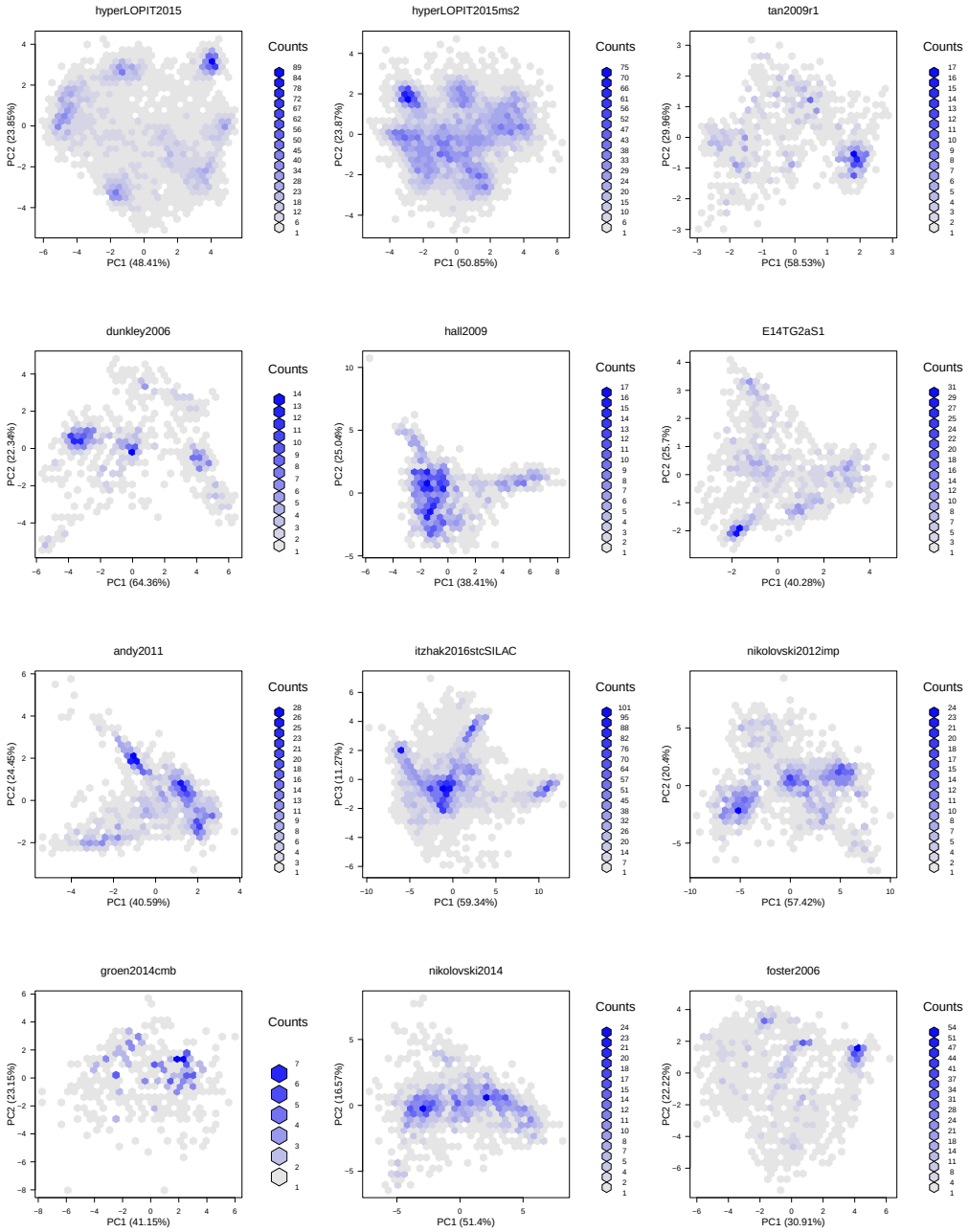

Figure 6: Density PCA plots for the 12 experiments used in this study. The experiments are ordered according to the median average between cluster distance (see figure 4).

Acknowledgements

L.M.B is supported by a Wellcome Trust Technology Development Grant (WTxxxxx). L.G. is supported by the BBSRC Strategic Longer and Larger grant (Award BB/L002817/1)

6 Session information

The software and versions used to produce this document are summarised below. In particular, recent versions of `pRoloc` and `pRolocdata` [Gatto et al., 2014b], which require versions 1.13.9 and 1.11.2 or later, respectively.

- R version 3.3.1 Patched (2016-08-02 r71022), `x86_64-pc-linux-gnu`
- Base packages: base, datasets, graphics, grDevices, methods, parallel, stats, stats4, utils
- Other packages: annotate 1.51.0, AnnotationDbi 1.35.4, Biobase 2.33.0, BiocGenerics 0.19.2, BiocParallel 1.7.8, cluster 2.0.4, hexbin 1.27.1, IRanges 2.7.12, MLInterfaces 1.53.1, MSnbase 1.99.0, mzR 2.7.3, pRoloc 1.13.11, pRolocdata 1.11.4, ProtGenerics 1.5.1, Rcpp 0.12.6, S4Vectors 0.11.10, XML 3.98-1.4, xtable 1.8-2
- Loaded via a namespace (and not attached): affy 1.51.0, affyio 1.43.0, assertthat 0.1, base64enc 0.1-3, BiocInstaller 1.23.6, biomaRt 2.29.2, bitops 1.0-6, car 2.1-3, caret 6.0-71, class 7.3-14, codetools 0.2-14, colorspace 1.2-6, DBI 0.5, DEoptimR 1.0-6, digest 0.6.10, diptest 0.75-7, doParallel 1.0.10, dplyr 0.5.0, e1071 1.6-7, evaluate 0.9, flexmix 2.3-13, FNN 1.1, foreach 1.4.3, formatR 1.4, fpc 2.1-10, gbm 2.1.1, gdata 2.17.0, genefilter 1.55.2, ggplot2 2.1.0, ggvis 0.4.3, grid 3.3.1, gtable 0.2.0, gtools 3.5.0, highr 0.6, htmltools 0.3.5, htmlwidgets 0.7, httpuv 1.3.3, hwriter 1.3.2, impute 1.47.0, iterators 1.0.8, jsonlite 1.0, kernlab 0.9-24, knitr 1.14, lattice 0.20-33, limma 3.29.17, lme4 1.1-12, lpSolve 5.6.13, magrittr 1.5, MALDIquant 1.15, MASS 7.3-45, Matrix 1.2-6, MatrixModels 0.4-1, mclust 5.2, mgcv 1.8-13, mime 0.5, minqa 1.2.4, mlbench 2.1-1, modeltools 0.2-21, munsell 0.4.3, mvtnorm 1.0-5, mzID 1.11.2, nlme 3.1-128, nloptr 1.0.4, nnet 7.3-12, pbkrtest 0.4-6,

pcaMethods 1.65.0, pls 2.5-0, plyr 1.8.4, prabclus 2.2-6, preprocessCore 1.35.0, proxy 0.4-16, quantreg 5.26, R6 2.1.2, randomForest 4.6-12, RColorBrewer 1.1-2, RCurl 1.95-4.8, rda 1.0.2-2, reshape2 1.4.1, robustbase 0.92-6, rpart 4.1-10, RSQLite 1.0.0, sampling 2.7, scales 0.4.0, sfsmisc 1.1-0, shiny 0.13.2, SparseM 1.7, splines 3.3.1, stringi 1.1.1, stringr 1.0.0, survival 2.39-5, threejs 0.2.2, tibble 1.1, tools 3.3.1, trimcluster 0.1-2, vsn 3.41.0, zlibbioc 1.19.0

References

L M Breckels, L Gatto, A Christoforou, A J Groen, K S Lilley, and M W Trotter. The effect of organelle discovery upon sub-cellular protein localisation. *J Proteomics*, 88:129–40, Aug 2013. doi: 10.1016/j.jprot.2013.02.019.

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.

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, 2016. doi: 10.1038/ncomms9992.

Tom P. J. Dunkley, Svenja Hester, Ian P. Shadforth, John Runions, Thilo Weimar, Sally L. Hanton, Julian L. Griffin, Conrad Bessant, Federica Brandizzi, Chris Hawes, Rod B. Watson, Paul Dupree, and Kathryn S. Lilley. Mapping the arabidopsis organelle proteome. *Proc Natl Acad Sci USA*, 103(17):6518–6523, Apr 2006. doi: 10.1073/pnas.0506958103. URL `http://dx.doi.org/10.1073/pnas.0506958103`.

L J Foster, C L de Hoog, Y Zhang, Y Zhang, X Xie, V K. Mootha, and M Mann. A mammalian organelle map by protein correlation profiling. *Cell*, 125(1):187–199, Apr 2006. doi: 10.1016/j.cell.2006.03.022.

L Gatto, J A Vizcaíno, H Hermjakob, W Huber, and K S Lilley. Organelle proteomics experimental designs and analysis. *Proteomics*, 10(22):3957–69, Nov 2010. doi: 10.1002/pmic.201000244.

L Gatto, L M Breckels, T Burger, D J Nightingale, A J Groen, C Campbell, N Nikolovski, C M Mulvey, A Christoforou, M Ferro, and K S Lilley. A foundation for reliable spatial proteomics data analysis. *Mol Cell Proteomics*, 13(8):1937–52, Aug 2014a. doi: 10.1074/mcp.M113.036350.

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*, 30(9):1322–4, May 2014b. doi: 10.1093/bioinformatics/btu013.

A J Groen, G Sancho-Andrs, L M Breckels, L Gatto, F Aniento, and K S Lilley. Identification of trans-golgi network proteins in arabidopsis thaliana root tissue. *J Proteome Res*, 13(2):763–76, Feb 2014. doi: 10.1021/pr4008464.

S L Hall, S Hester, J L Griffin, K S Lilley, and A P Jackson. The organelle proteome of the dt40 lymphocyte cell line. *Mol Cell Proteomics*, 8(6):1295–1305, Jun 2009. doi: 10.1074/mcp.M800394-MCP200. URL `http://dx.doi.org/10.1074/mcp.M800394-MCP200`.

D N Itzhak, S Tyanova, J Cox, and G H Borner. Global, quantitative and dynamic mapping of protein subcellular localization. *Elife*, 5, 2016. doi: 10.7554/eLife.16950.

N Nikolovski, P V Shliaha, L Gatto, P Dupree, and K S Lilley. Label free protein quantification for plant golgi protein localisation and abundance. *Plant Physiol*, Aug 2014. doi: 10.1104/pp.114.245589.