Computational analyses of high-throughput spatial proteomics data

L. Gatto, L. Breckels and K.S. Lilley
University of Cambridge

18 July 2013

Spatial/organelle proteomics - Why

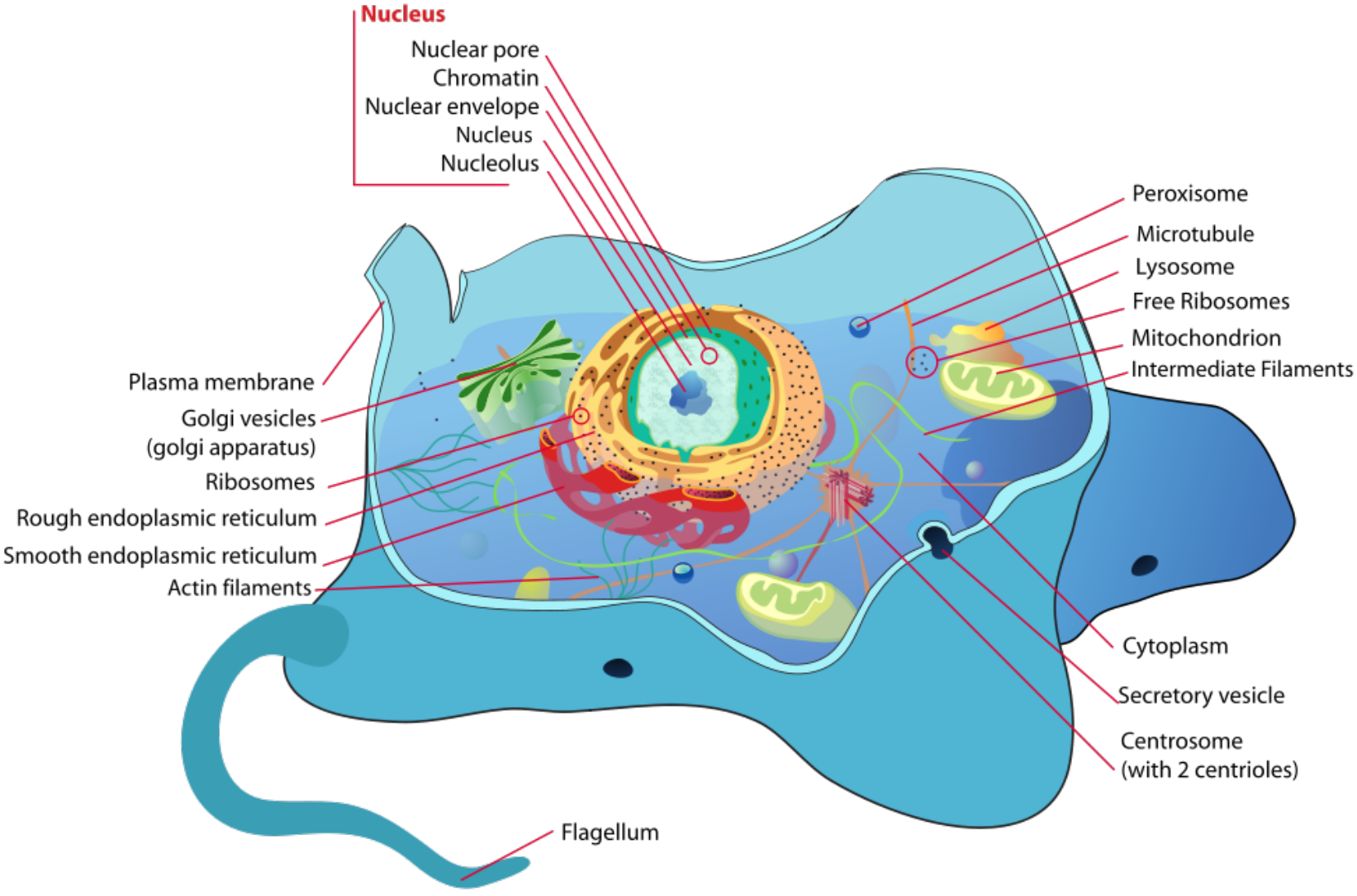
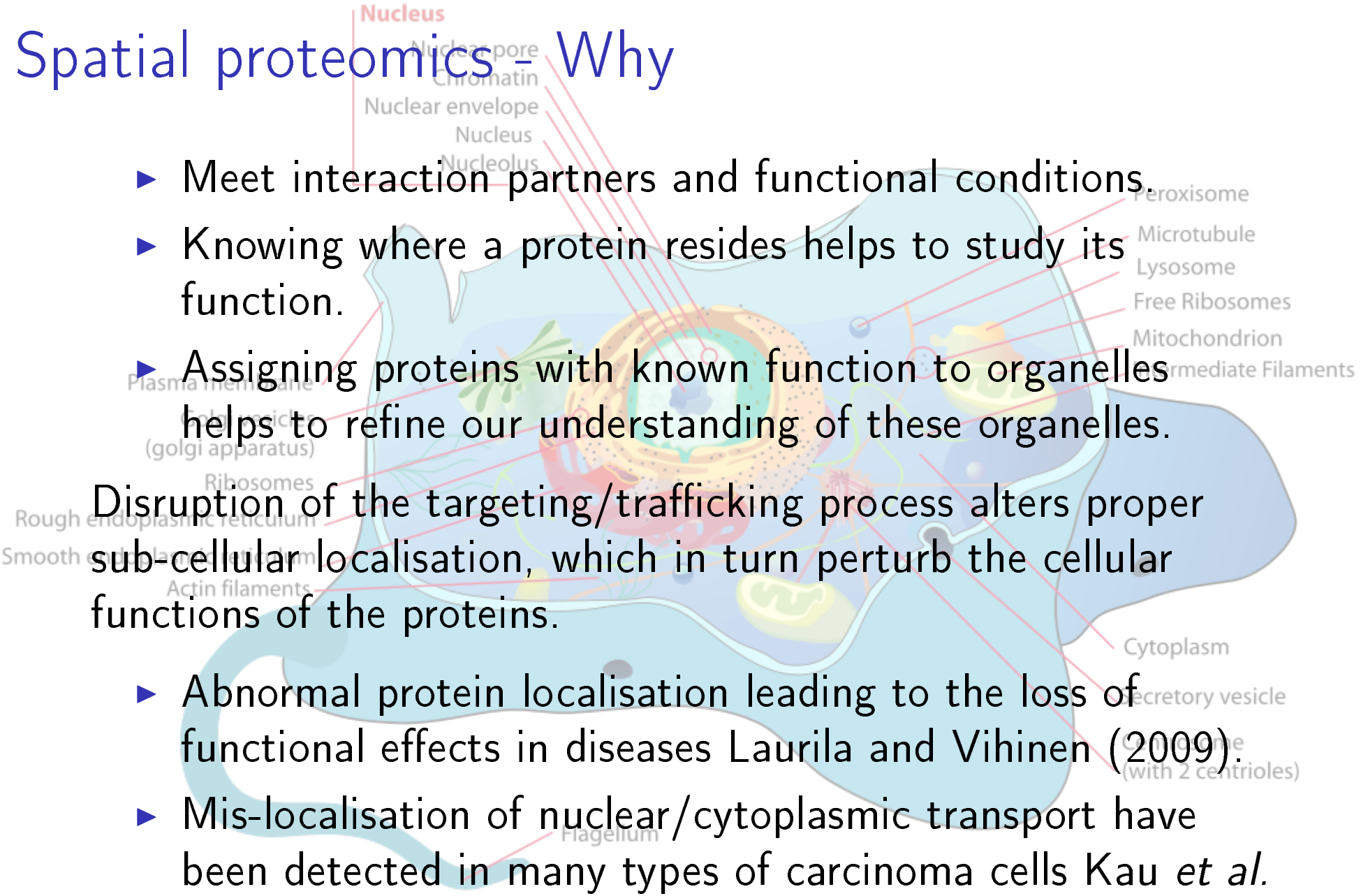


Image from Wikipedia http://en.wikipedia.org/wiki/Cell_(biology).

Spatial proteomics - Why

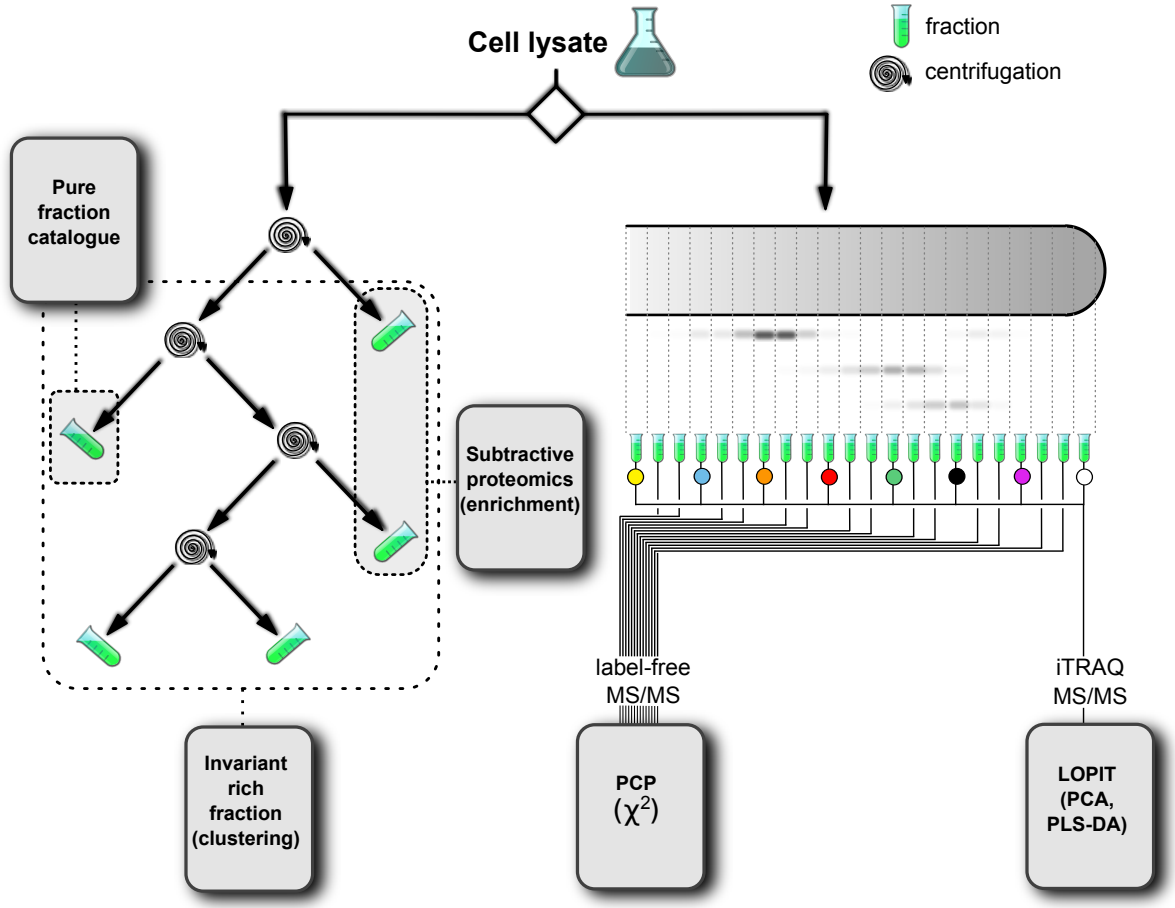


- Meet interaction partners and functional conditions.
- Knowing where a protein resides helps to study its function.
- Assigning proteins with known function to organelles helps to refine our understanding of these organelles.

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).
- Mis-localisation of nuclear/cytoplasmic transport have been detected in many types of carcinoma cells Kau *et al.* (2004).

Spatial proteomics - How, experimentally



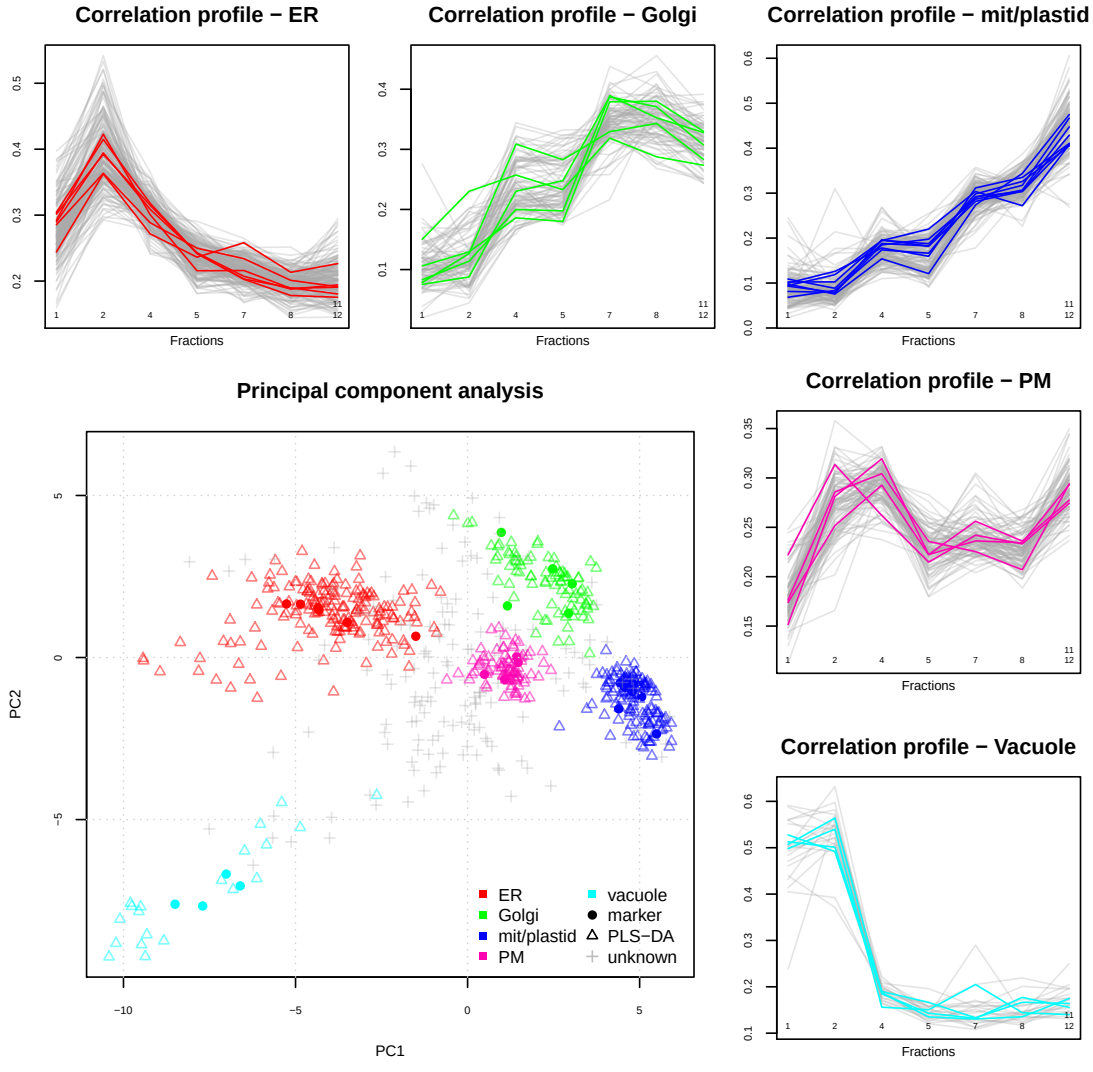
From Gatto *et al.* (2010).

Computationally

Stating the problem from a computational point of view.

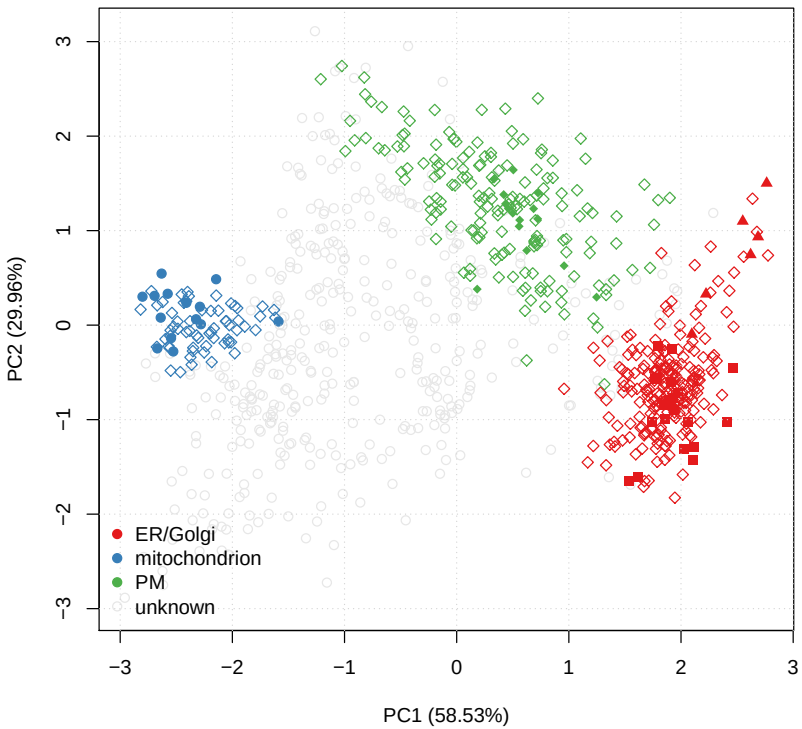
	Fraction_1	Fraction_2	...	Fraction_m	markers
p_1	$q_{1,1}$	$q_{1,2}$	...	$q_{1,\,m}$	loc_1
p_2	$q_{2,1}$	$q_{2,2}$	...	$q_{2,\,m}$	loc_2
p_3	$q_{3,1}$	$q_{3,2}$	...	$q_{3,\,m}$	
p_4	$q_{4,1}$	$q_{4,2}$	...	$q_{4,\,m}$	loc_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
p_i	$q_{i,1}$	$q_{i,2}$	...	$q_{i,\,m}$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
p_n	$q_{n,1}$	$q_{n,2}$	...	$q_{n,\,m}$	loc_k

Visually



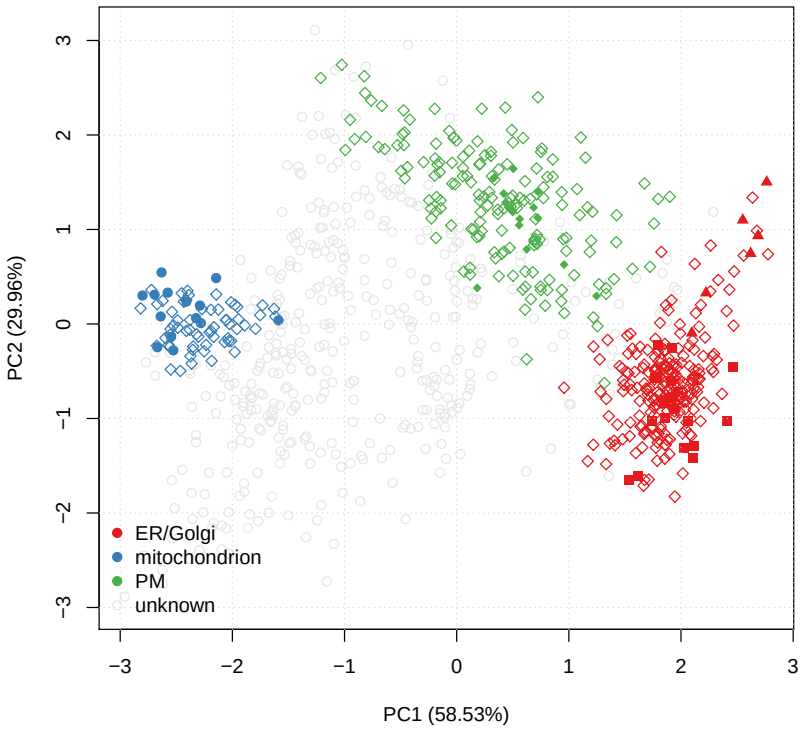
From Gatto *et al.* (2010), data from Dunkley *et al.* (2006)

Then

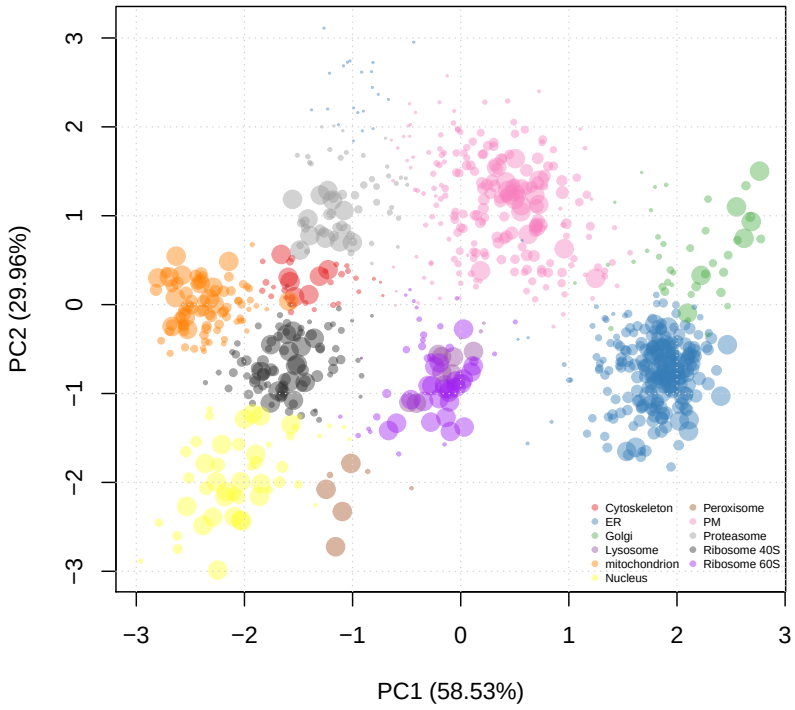


Data as presented in Tan *et al.* (2009)

Then and now



Data as presented in Tan *et al.* (2009)

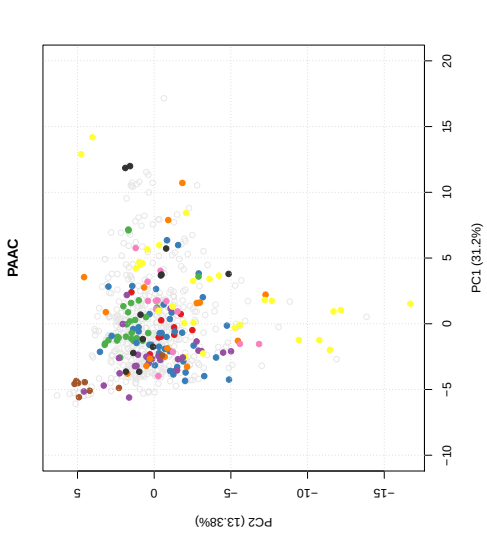


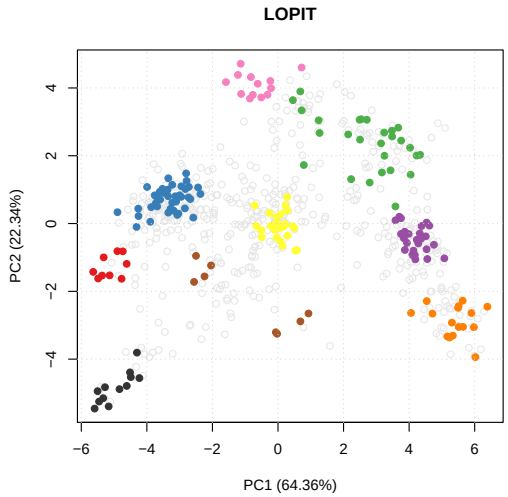
Augmented marker set using novelty detection from (Breckels *et al.*, 2013) and class-weighted svm with classifier posterior probabilities.

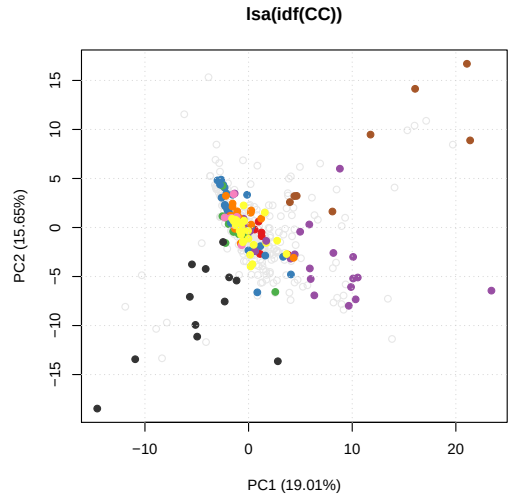
Dry approaches

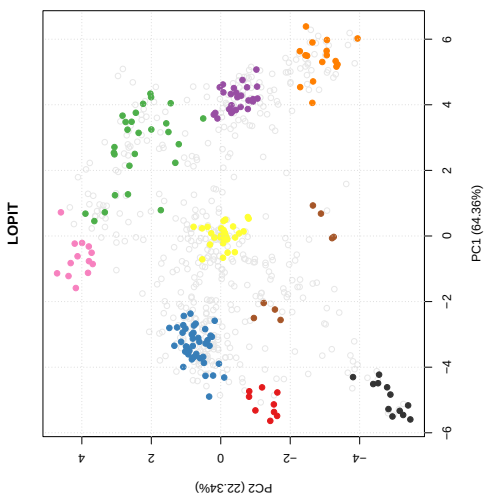
Using **sorting signals** or protein domains, **gene ontology** terms, **sequence features** (Chou, 2001) or combinations of the these.

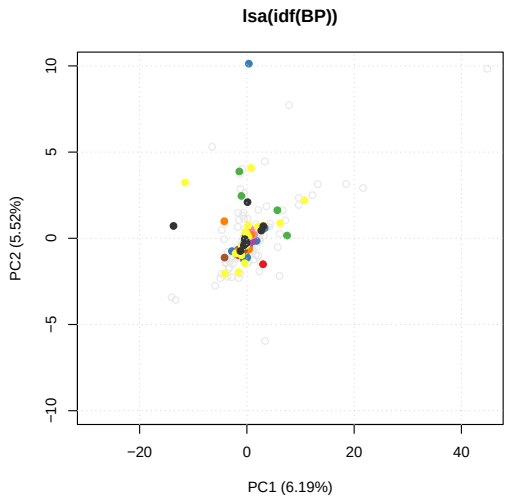
- **free/cheap** vs. expensive
- abundant (full proteome, 25000 entries) vs. **targeted** (500 – 2000 proteins)
- *low* vs. *high* **quality**

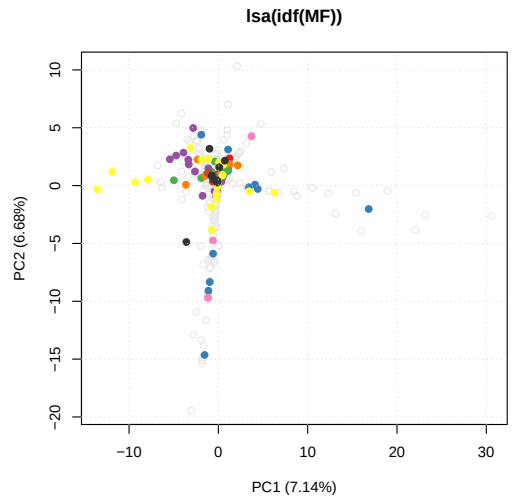

PAAC
PC1 (31.2%)
PC2 (13.38%)


LOPIT
PC2 (22.34%)
PC1 (64.36%)


lsa(idf(CC))
PC2 (15.65%)
PC1 (19.01%)


LOPIT
PC1 (64.36%)
PC2 (22.34%)


lsa(idf(BP))
PC2 (5.52%)
PC1 (6.19%)


lsa(idf(MF))
PC2 (6.68%)
PC1 (7.14%)

Dry approaches

Using **sorting signals** or protein domains, **gene ontology** terms, **sequence features** (Chou, 2001) or combinations of the these.

- **free/cheap** vs. expensive
- abundant (full proteome, 25000 entries) vs. **targeted** (500 – 2000 proteins)
- *low* vs. *high* **quality**

- **static** vs. **dynamic**

Getting the best out of each data source

- **Data fusion**: good for (high quality) exp data only (Trotter *et al.*, 2010) but highly detrimental when fusing high and low quality data.
- A **Weight Adjusted Voting classification Ensemble** (Kim *et al.*, 2011): Iteratively assigns weights to each classifier (i.e. each source of information) in the ensemble and another weight vector for all instances

- LOPIT (n x 16 matrix)

	M1F1A	M1F4A	M2F8B	M2F11B
AT1G03860	0.112143	0.192714	0.3215	0.4205
AT1G07810	0.275000	0.276000	0.2385	0.2025
AT1G08660	0.038800	0.252200	0.3374	0.2802

- Gene Ontology - Molecular Function ($n \times 293$)
- Gene Ontology - Cellular Component ($n \times 115$)

	GO:0005783	GO:0005739	GO:0010008	GO:0033178
AT1G03860	0	1	0	0
AT1G07810	1	0	0	0
AT1G08660	0	0	0	0

- Amino acid sequence – Pseudo amino acid code ($n \times 50$)

	PAAC1	PAAC2	PAAC49	PAAC50
AT1G03860	7.87424	4.921400	0.02474136	0.02536978
AT1G07810	21.61686	11.742490	0.02435662	0.02451319
AT1G08660	6.51358	8.443529	0.02651188	0.02496825

Classifier weights

	p_weight
LOPIT	0.46988507
PAAC	0.09459885
GO.CC	0.33377615
GO.MF	0.10173993

Example weights

	q_weight
AT1G03860	0.008799044
AT1G07810	0.008799044
AT1G08660	0.000000000
AT1G09210	0.012049173
...	
AT5G66680	0.000000000
AT5G67500	0.000000000

Accuracy

	LOPIT	PAAC	GO.CC	GO.MF	MAJ.VOTE	WAVE
Dunkley (2006)	0.945	0.459	0.824	0.482	0.915	0.934
Tan (2009)	0.885	0.344	0.550	0.402	0.785	0.880
Andy (HEK293)	0.827	0.300	0.723	0.325	0.712	0.815

Software

Infrastructure: MSnbase, ML: pRoloc and data: pRolocdata.

References

Breckels, L. M. et al. (2013). The effect of organelle discovery upon sub-cellular protein localisation. *J Proteomics*.

Chou, K. C. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins*, 43(3), 246–55.

Dunkley, T. et al. (2006). Mapping the arabidopsis organelle proteome. *Proc Natl Acad Sci USA*, 103(17), 6518–6523.

Gatto, L. et al. (2010). Organelle proteomics experimental designs and analysis. *Proteomics*.

Kau, T. R. et al. (2004). Nuclear transport and cancer: from mechanism to intervention. *Nat Rev Cancer*, 4(2), 106–17.

Kim, H. et al. (2011). A weight-adjusted voting algorithm for ensembles of classifiers. *Journal of the Korean Statistical Society*, 40(4), 437 – 449.

Laurila, K. et al. (2009). Prediction of disease-related mutations affecting protein localization. *BMC Genomics*, 10, 122.

Tan, D. J. et al. (2009). Mapping organelle proteins and protein complexes in drosophila melanogaster. *J Proteome Res*, 8(6), 2667–78.

Trotter, M. et al. (2010). Improved sub-cellular resolution via simultaneous analysis of organelle proteomics data across varied experimental conditions. *PROTEOMICS*, 10(23), 4213–4219.

Acknowledgements

- Kathryn Lilley and Campridge Centre for Proteomics
- Lisa Breckels (phenoDisco, data fusion)
- Those that produce the actual data
- Funding: PRIME-XS FP7 and BBSRC

Thank you for your attention.