

Introduction to MS-based proteomics and Bioconductor infrastructure

Laurent Gatto

`lg390@cam.ac.uk` – `@lgatto`

`http://cpu.sysbiol.cam.ac.uk`

CSAMA – 17 June 2015

Outline

Proteomics and MS data

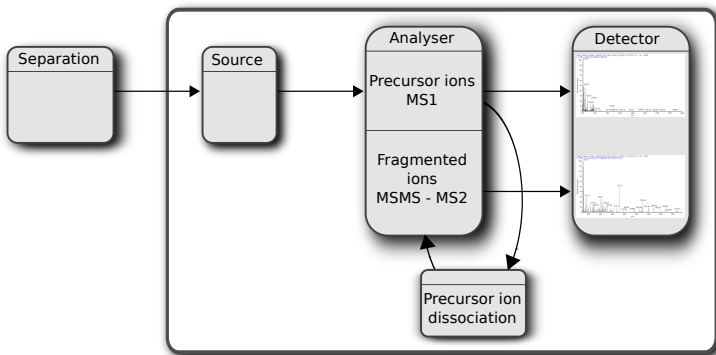
Bioconductor infrastructure

Examples

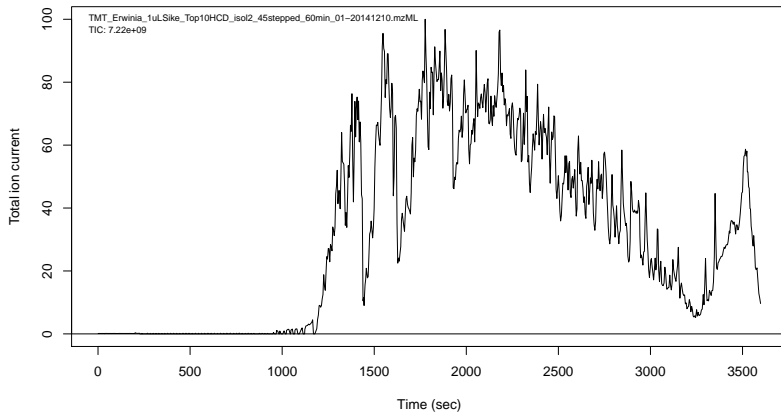
Ranges infrastructure

Application: spatial proteomics

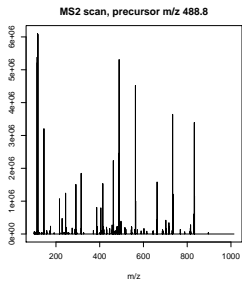
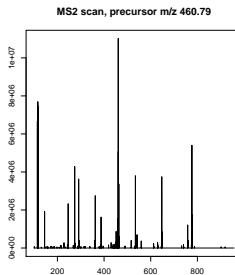
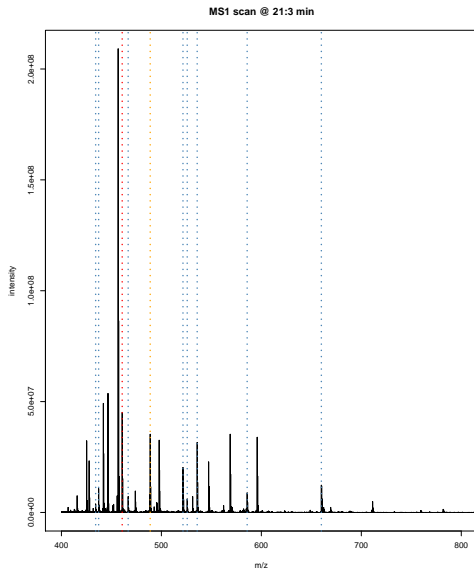
Mass-spectrometry – LC-**MS**/MS



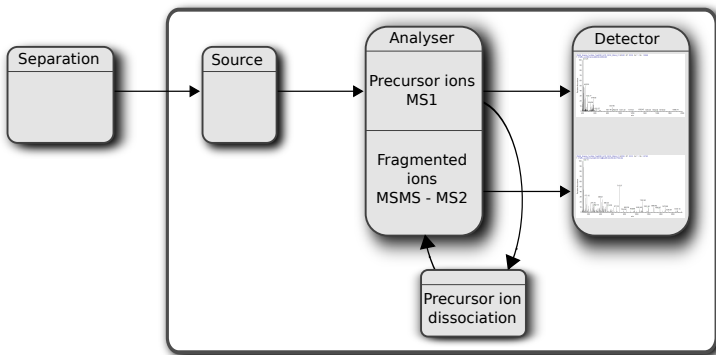
Chromatogram: total intensity over time



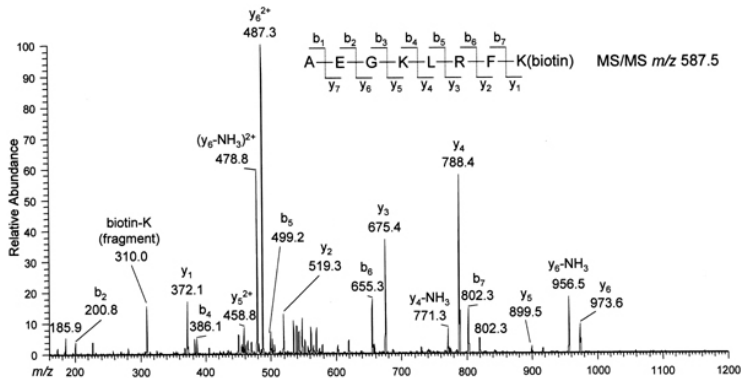
MS1 (and MS2) spectra



Mass-spectrometry – LC-**MS**/MS



Fragmentation



Credit abrg.org

```
cid <- calculateFragments("AEGKLRFK",  
                           type=c("b", "y"), z=2)
```

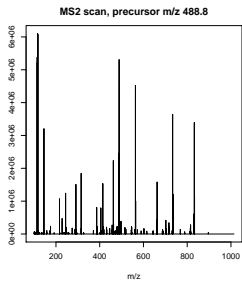
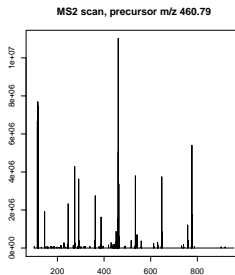
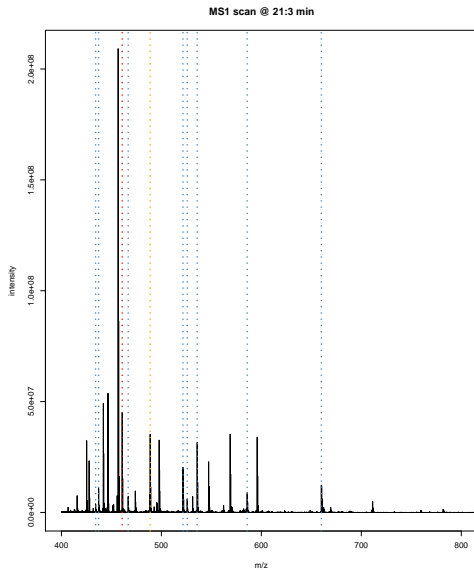
```
## Modifications used: C=160.030649
```

```
ht(cid, n = 3)
```

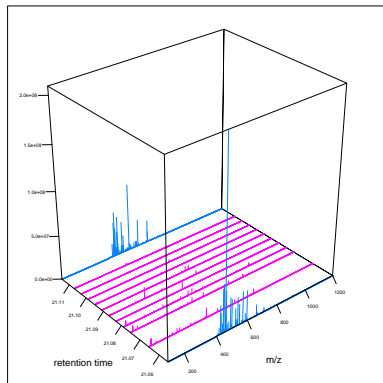
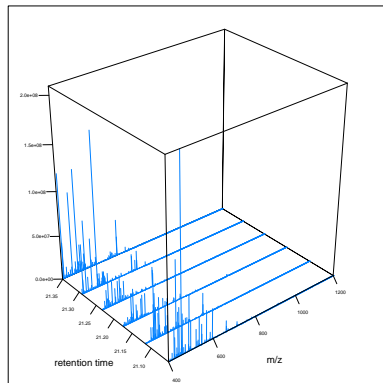
```
##           mz ion type pos z seq  
## 1  36.52583  b1    b   1 2  A  
## 2 101.04713  b2    b   2 2  AE  
## 3 129.55786  b3    b   3 2  AEG  
## ...
```

```
##           mz ion type pos z      seq  
## 31 357.7185 y6*   y*   6 2    GKLRFK  
## 32 422.2398 y7*   y*   7 2    EGKLRFK  
## 33 457.7583 y8*   y*   8 2  AEGKLRFK
```


MS1 and MS2 spectra



MS1 and MS2 spectra



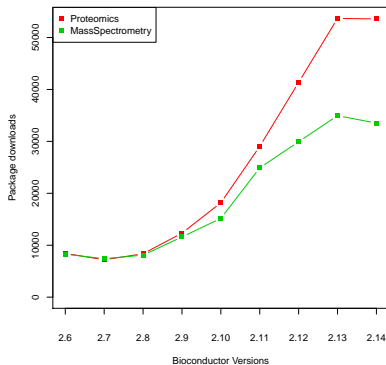
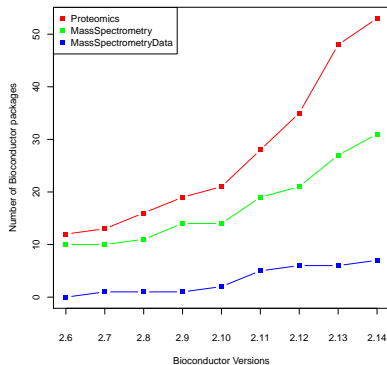
Proteomics data

- ▶ raw data:
MS1 and MS2 over retention time
 - ▶ identification:
MS2
 - ▶ quantitation:
MS1 or MS2
- ▶ protein database
(to match MS2 spectra against)

	Status	package
Raw (mz*ML)	✓	mzR
mzTab	✓	MSnbase
mgf	✓	MSnbase
mzIdentML	✓	mzID, mzR
mzQuantML		(?mzR)

Bioconductor infrastructure

biocViews: Proteomics, MassSpectrometry



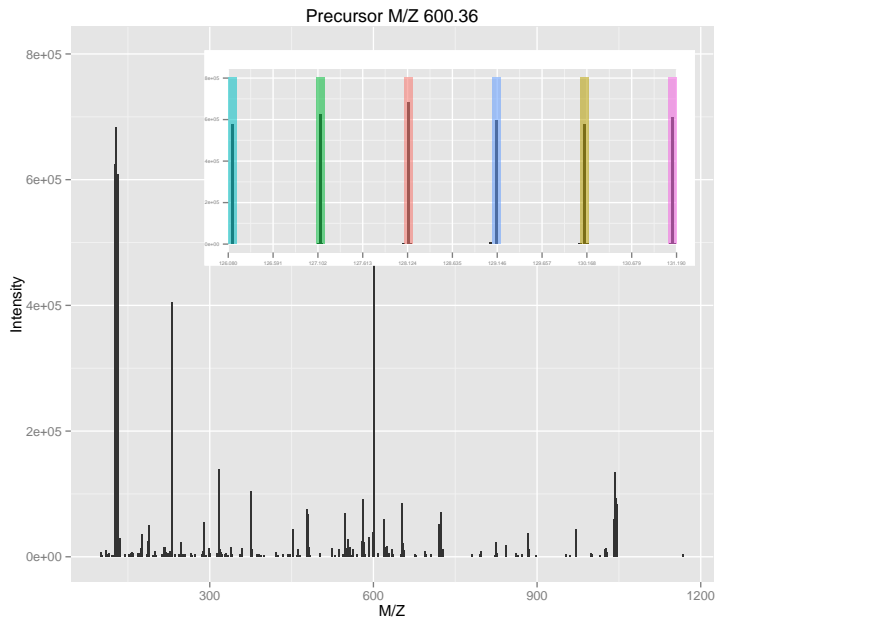
Learning from Bioconductor

genomics	proteomics
-----+-----	
eSet (past?)	*MSnSet (present)
Ranges (present)	*Pbase et al. (future)
	PPI (present)
	*localisation (present)

Example

```
library("MSnbase")  
rx <- readMSData("rawdata.mzML")  
rx <- addIdentificationData(rx, "identification.mzid")  
rx <- rx[!is.na(fData(rx)$pepseq)]  
plot(rx[[10]], reporters = TMT6, full=TRUE)
```

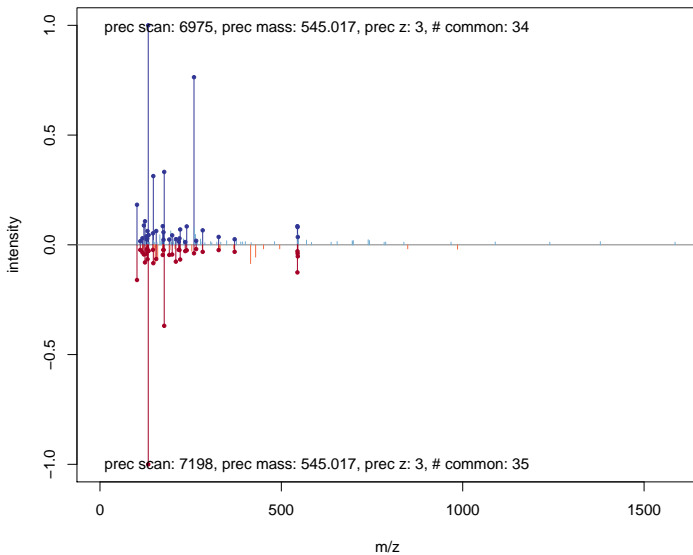
Example



Example

```
library("MSnbase")
rx <- readMSData(f, centroided = TRUE)
rx <- addIdentificationData(rx, g)
rx <- rx[!is.na(fData(rx)$pepseq)]
plot(rx[[10]], reporters = TMT6, full=TRUE)
plot(rx[[4730]], rx[[4929]])
```


Example

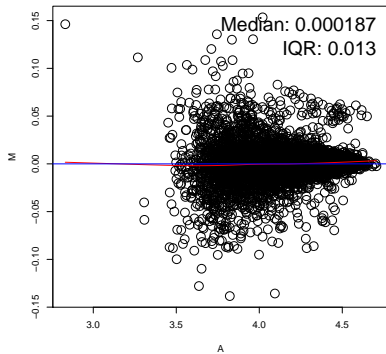
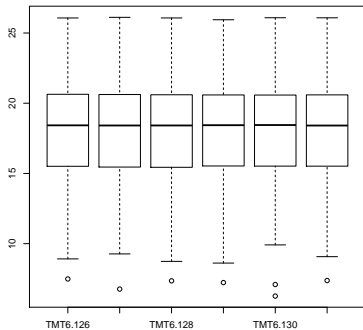


Example

```
library("MSnbase")
rx <- readMSData(f, centroided = TRUE)
rx <- addIdentificationData(rx, g)
rx <- rx[!is.na(fData(rx)$pepseq)]
plot(rx[[10]], reporters = TMT6, full=TRUE)
plot(rx[[4730]], rx[[4929]])

qt <- quantify(rx, reporters = TMT6)
## qt <- readMSnSet("quantdata.csv", cols = 5:11)
nqt <- normalise(qt, method = "vsn")
boxplot(exprs(nqt))
MAplot(nqt[, 1:2])
```

Example



More

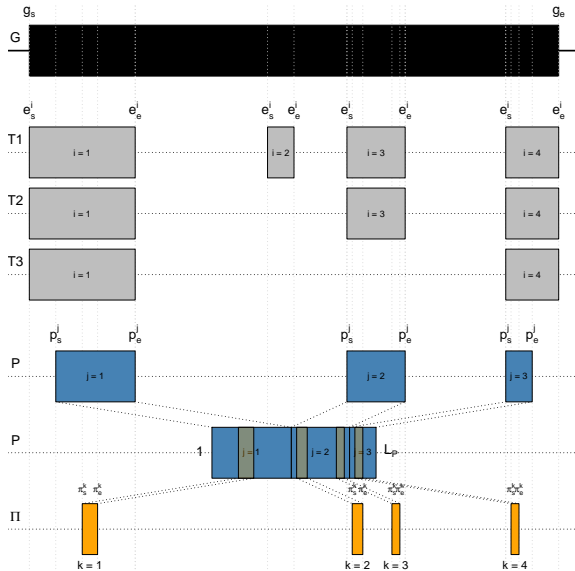
- ▶ RforProteomics package

```
library("RforProteomics")  
RforProteomics()  
RProtVis()  
citation(package = "RforProteomics")
```

- ▶ Proteomics workflow on the Bioc site
- ▶ Lab on Friday

- ▶ **protein database**
- ▶ raw data
 - ▶ quantitation
 - ▶ identification

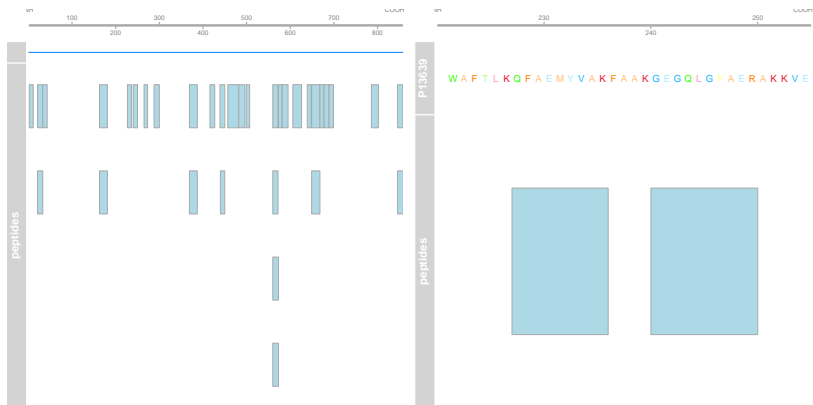
Ranges infrastructure



Pbase package

```
library("Pbase")  
p <- Proteins("uniprot.fasta")  
p <- addIdentificationData(p, "identification.mzid")  
aa(p) ## peptides sequences as a AAStringSet  
pranges(p) ## peptide ranges as IRangesList  
i <- which(acols(p)[, "EntryName"] == "EF2_HUMAN")  
plot(p[i])  
plot(p[i], from = 155, to = 185)
```

Along protein coordinates



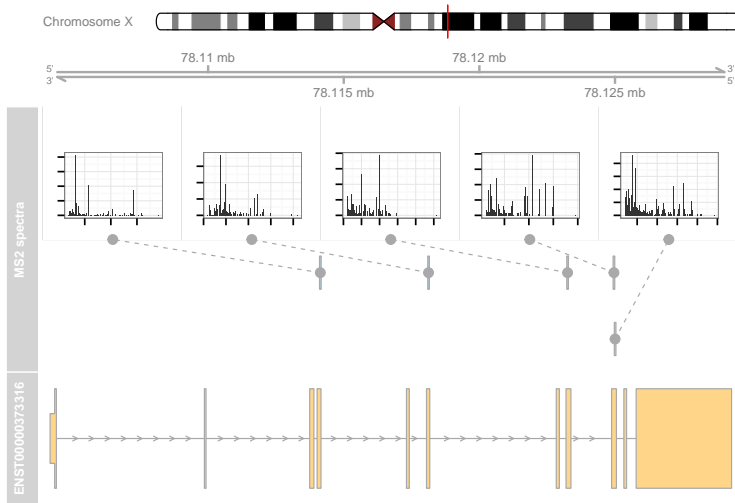
Along genome coordinates

... using transcript models as GRangesList and Gviz for plotting.



From the **Pbase** mapping vignette.

Along genome coordinates (with raw data)



From the **Pbase** mapping vignette.

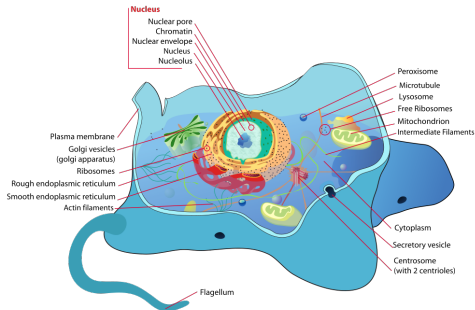
With RNA-Seq reads



From <https://github.com/ComputationalProteomicsUnit/Intro-Integ-Omics-Prot>

Spatial proteomics

- ▶ The cellular sub-division allows cells to establish a range of distinct microenvironments, each favouring different biochemical reactions and interactions and, therefore, allowing each compartment to fulfil a particular functional role.
- ▶ Localisation and sequestration of proteins within subcellular niches is a fundamental mechanism for the post-translational regulation of protein function.



Spatial proteomics is the systematic study of protein localisations.

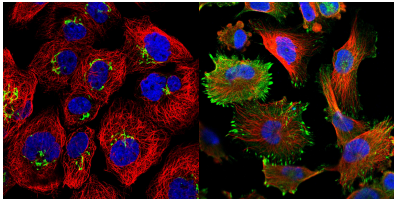


Figure : Immunofluorescence: ZFPL1, Golgi (left) and FHL2, mainly localized to actin filaments and focal adhesion sites. Also detected in the nucleus (right). (from the Human Protein Atlas)

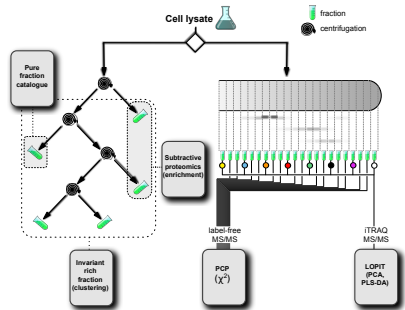


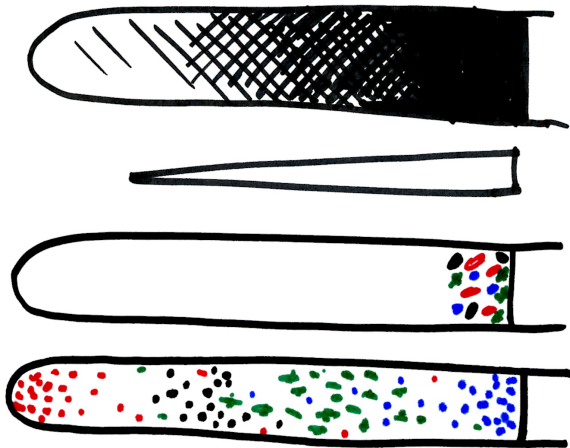
Figure : Mass spectrometry-based approaches based on density gradient subcellular fractionation.

Cell membrane lysis

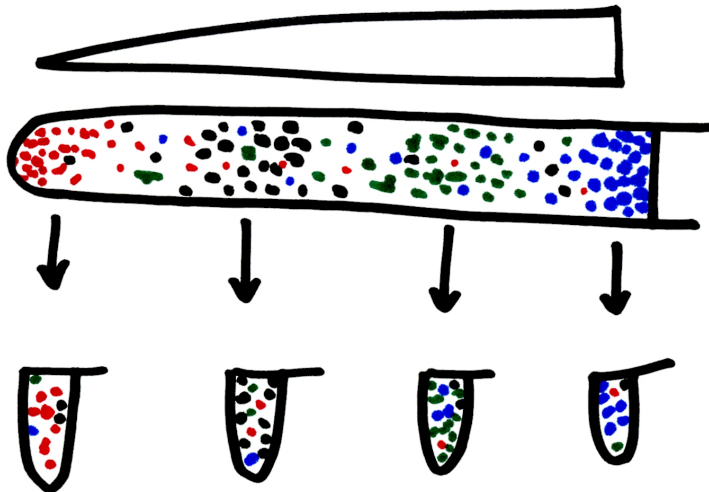
Mechanical or buffer-induced lysis of the plasma membrane with minimal disruption to intracellular organelles followed by subcellular fractionation.



Density gradient separation



Quantitation by LC-MSMS



Data

	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>loc</i> ₁
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> _k
⋮	⋮	⋮	⋮	⋮	⋮
p _n	q _{n,1}	q _{n,2}	...	q _{n, m}	unknown

Data analysis

MSnbase for data manipulation, **pRoloc** for clustering, classification and plotting, and **pRolocGUI** for interactive exploration.

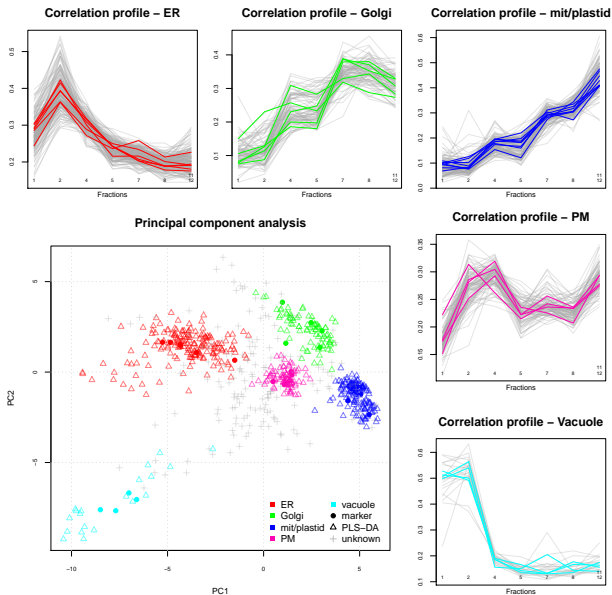


Figure : From Gatto et al. (2010), data from Dunkley et al. (2006).

2009 vs 2013

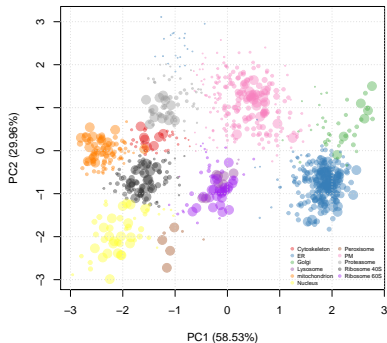
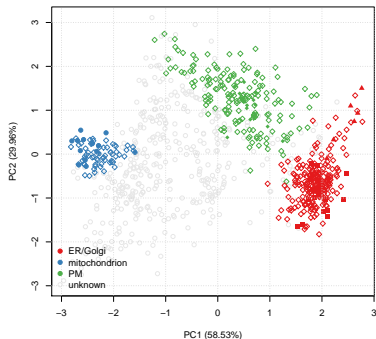
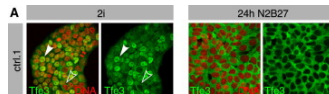
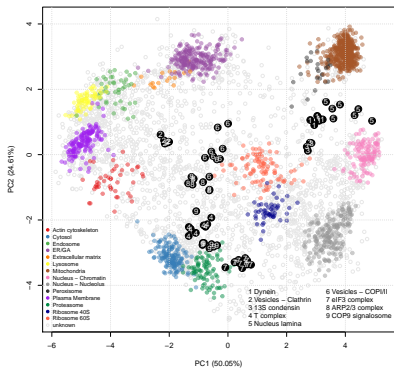
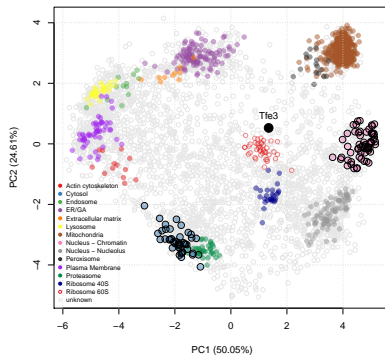


Figure : Semi-supervised approach Breckels et al. (2013). Data from Tan et al (2009).



From Betschinger *et al.* (2013)

Mouse ESC (E14TG2a) in serum LIF



Acknowledgement

- ▶ Lisa Breckels
- ▶ Sebastien Gibb
- ▶ Kathryn Lilley (CCP)

```
## R version 3.2.0 Patched (2015-04-22 r68234)
## Platform: x86_64-unknown-linux-gnu (64-bit)
## Running under: Ubuntu 14.04.2 LTS
##
## attached base packages:
## [1] stats      graphics  grDevices  utils      datasets  methods   base
##
## loaded via a namespace (and not attached):
## [1] Biobase_2.29.1          vsn_3.37.1
## [3] splines_3.2.0          foreach_1.4.2
## [5] Formula_1.2-1          affy_1.47.1
## [7] Pbase_0.9.0            highr_0.5
## [9] stats4_3.2.0           latticeExtra_0.6-26
## [11] BSgenome_1.37.1        Rsamtools_1.21.8
## [13] impute_1.43.0          RSQlite_1.0.0
## [15] lattice_0.20-31        biovizBase_1.17.1
## [17] limma_3.25.9           chron_2.3-45
## [19] digest_0.6.8           GenomicRanges_1.21.15
## [21] RColorBrewer_1.1-2     XVector_0.9.1
## [23] colorspace_1.2-6       preprocessCore_1.31.0
## [25] plyr_1.8.2             MALDIquant_1.12
## [27] XML_3.98-1.2           biomaRt_2.25.1
## [29] zlibbioc_1.15.0        scales_0.2.4
## [31] affyio_1.37.0          cleaver_1.7.0
## [33] BiocParallel_1.3.25    IRanges_2.3.11
## [35] ggplot2_1.0.1          SummarizedExperiment_0.1.5
## [37] GenomicFeatures_1.21.13 nnet_7.3-9
## [39] Gviz_1.13.2           BiocGenerics_0.15.2
## [41] proto_0.3-10          survival_2.38-1
## [43] magrittr_1.5           evaluate_0.7
## [45] doParallel_1.0.8       MASS_7.3-40
## [47] foreign_0.8-63        mzR_2.3.1
```