MS-based proteomics using Bioconductor

Laurent Gatto
lg390@cam.ac.uk

Cambridge Centre For Proteomics (CCP)
University of Cambridge

European Bioinformatics Institute (EBI)

1^{st} July 2011

Plan

Plan

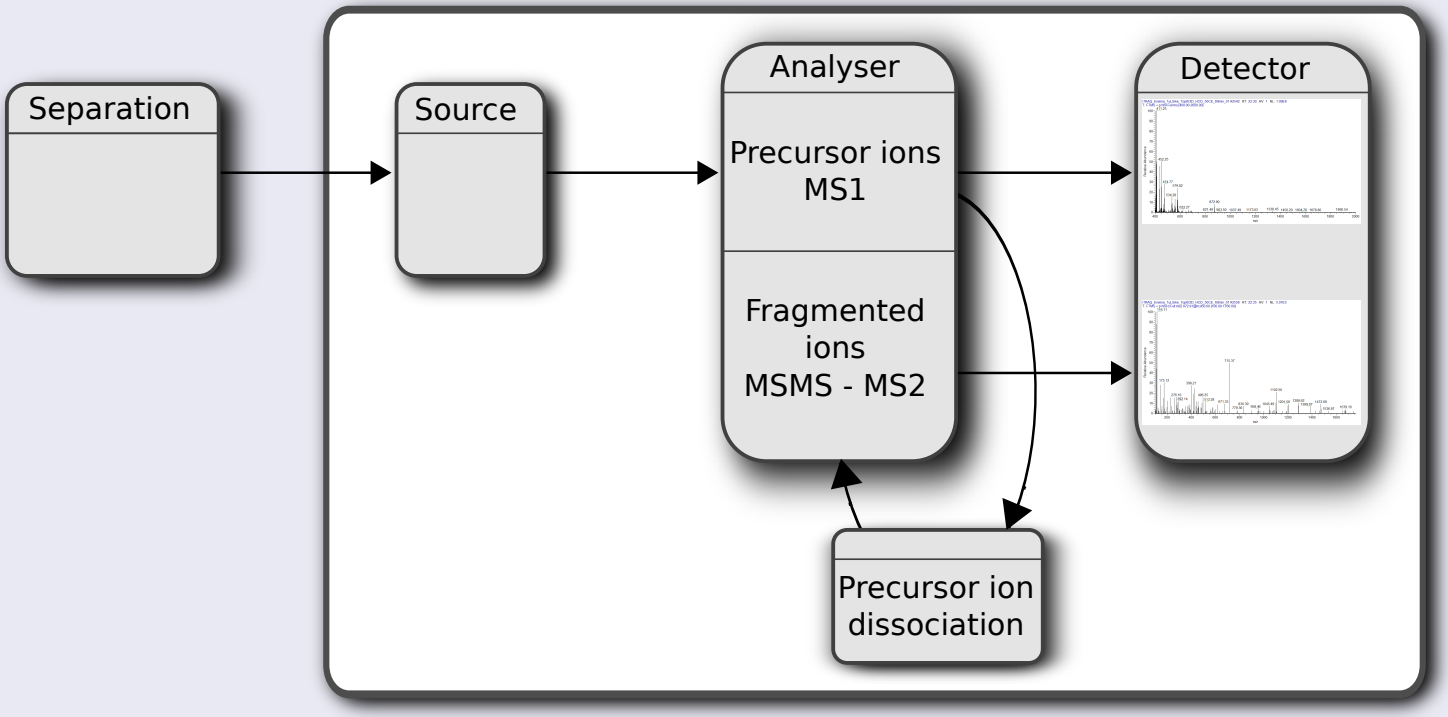
Motivation

- Many pieces of software are black boxes and *just* return values.
- Little/no solution to **explore** raw data and effect of processing/transformation.

Goals of `MSnbase`

- Apply the Bioconductor software model to MS-based proteomics
- Use robust and annotation rich data structure.
- Re-use algorithms readily available.
- Integration of genetic, genomic, proteomic, metabolomic data.

Schematic MS/MS workflow

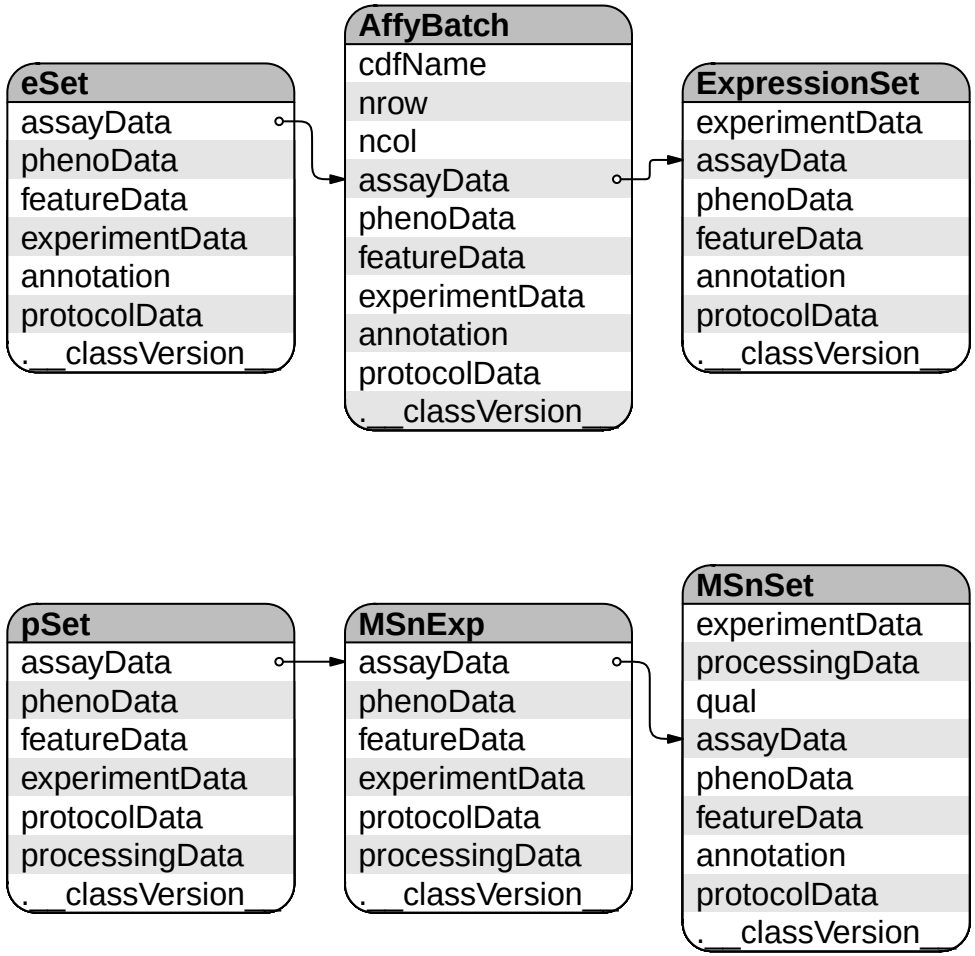


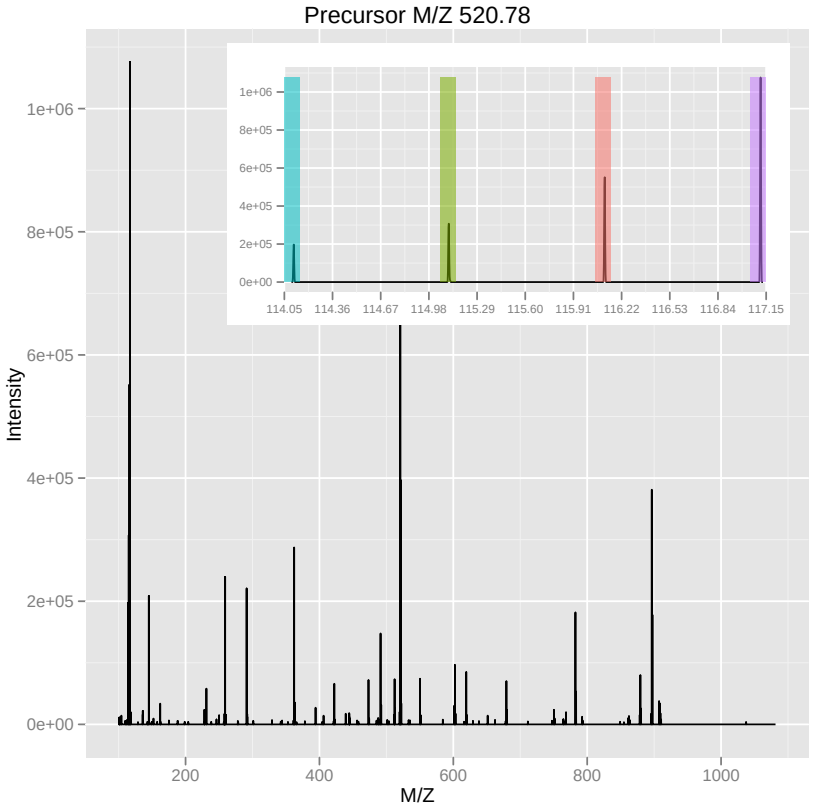
Plan

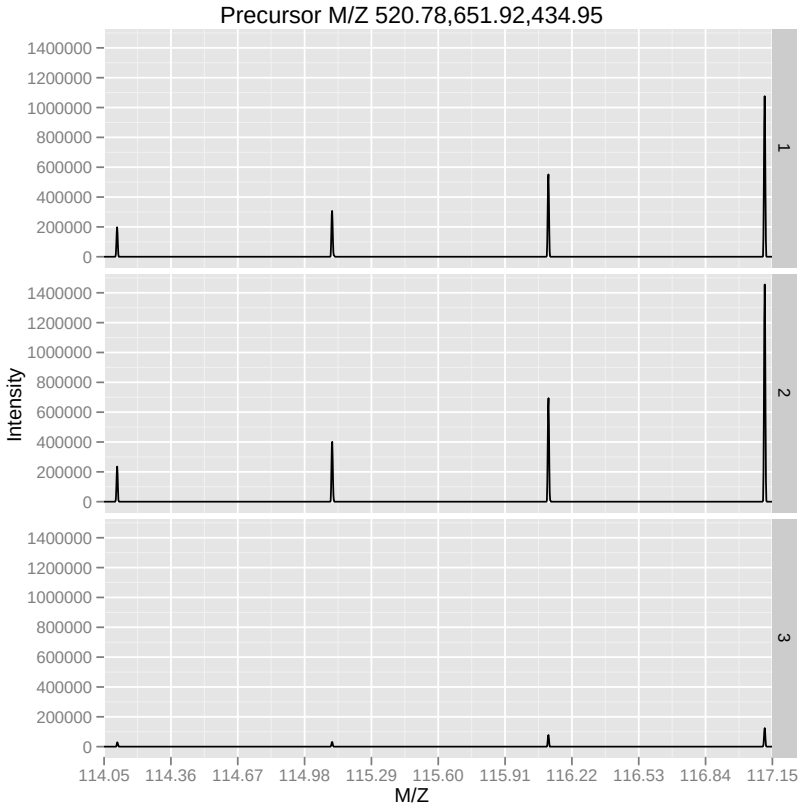
Classes

- `MSnExp` - MS(MS) experiment.
- `Spectrum`, `Spectrum1` and `Spectrum2` – mass spectra.
- `ReporterIons` defines reporter ions – `data(iTRAQ4)`.
- `MSnSet` – quantified expression.

- Additional meta-data in `MSnProcess` and `MIAPE`.


eSet
assayData
phenoData
featureData
experimentData
annotation
protocolData
.__classVersion__
AffyBatch
cdfName
nrow
ncol
assayData
phenoData
featureData
experimentData
annotation
protocolData
.__classVersion__
ExpressionSet
experimentData
assayData
phenoData
featureData
annotation
protocolData
.__classVersion__
pSet
assayData
phenoData
featureData
experimentData
protocolData
processingData
.__classVersion__
MSnExp
assayData
phenoData
featureData
experimentData
protocolData
processingData
.__classVersion__
MSnSet
experimentData
processingData
qual
assayData
phenoData
featureData
annotation
protocolData
.__classVersion__


Precursor M/Z 520.78
Intensity
M/Z
1e+06
8e+05
6e+05
4e+05
2e+05
0e+00
200
400
600
800
1000
114.05
114.36
114.67
114.98
115.29
115.60
115.91
116.22
116.53
116.84
117.15


Precursor M/Z 520.78,651.92,434.95
Intensity
M/Z
1400000
1200000
1000000
800000
600000
400000
200000
0
1
2
3
114.05
114.36
114.67
114.98
115.29
115.60
115.91
116.22
116.53
116.84
117.15

Plan

1. `readMzXMLData()` to create and `MSnExp` instance
2. plot() subset of `MSnExp` or `Spectrum`
3. Quality control (see later)
4. Processing: `removePeaks`, `bg.correct`
5. `quantify(MSnExp,ReporterIons)` to create an `MSnSet` instance
6. `purityCorrect(MSnSet,impurities)`
7. `normalise(MSnSet,"vsn")`
8. ...

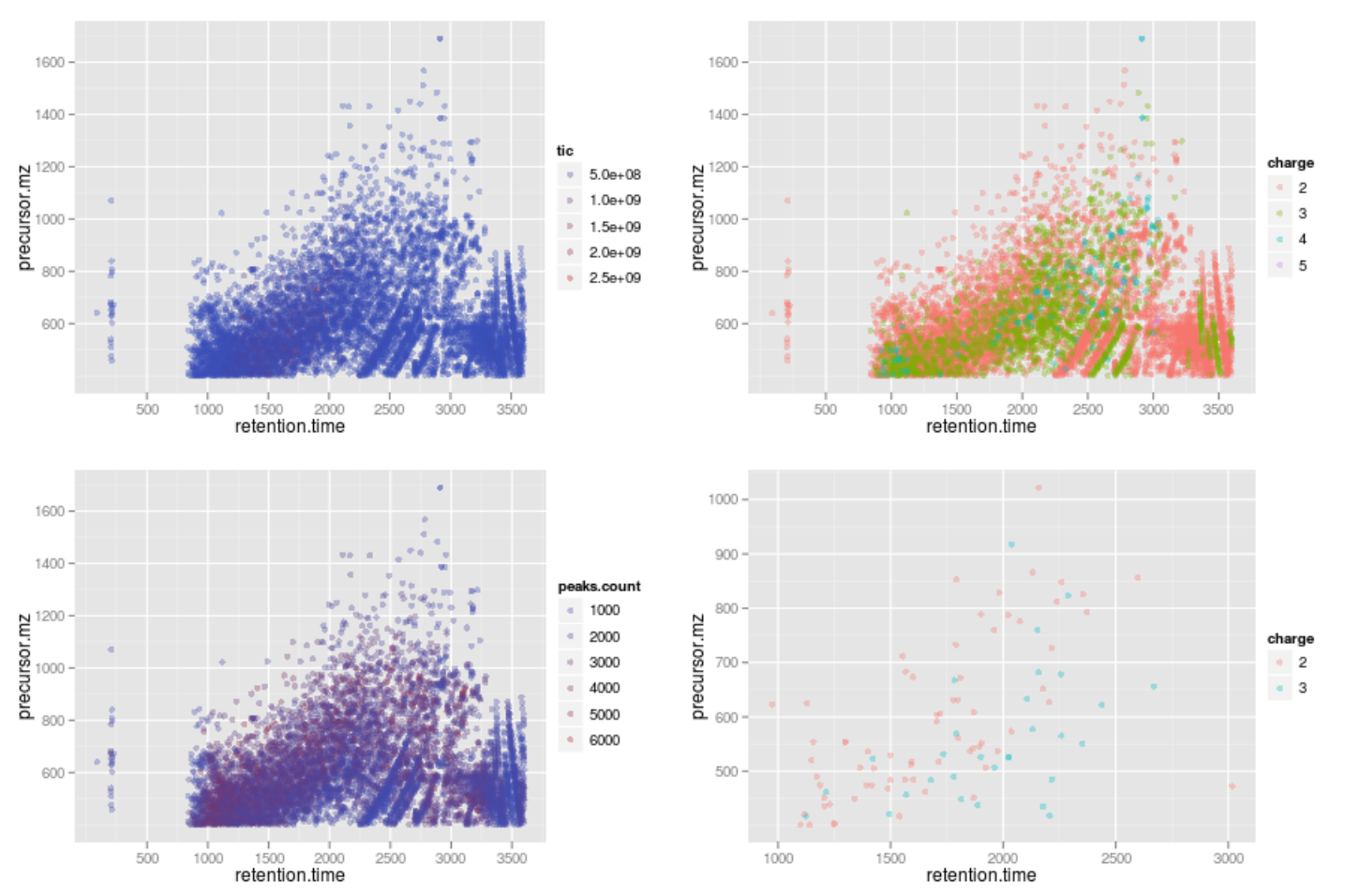
Number of times a precursor ion has been selected

Optimise MS parameters.

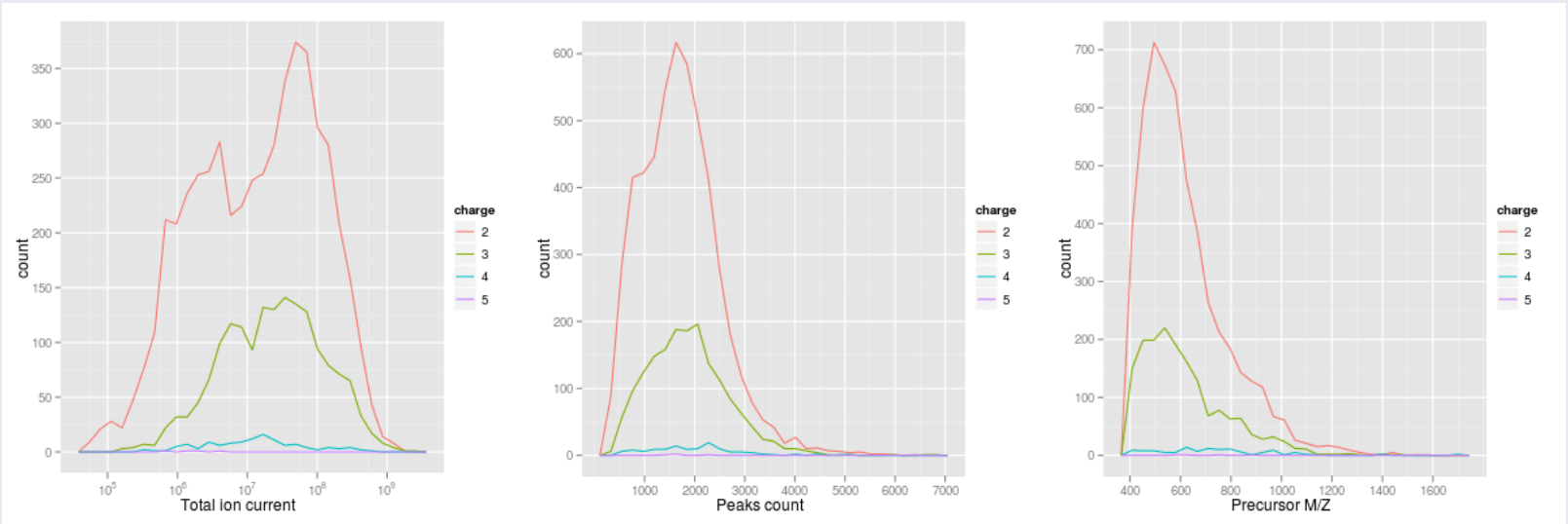
```
allPrecs <- precursorMz(raw)
number.selection <- c()
ms1scanNums <- ms1scan(raw)
for (mp in unique(allPrecs))
  number.selection <- c(number.selection,
                        length(unique(ms1scanNums[allPrecs==mp])))
names(number.selection) <- unique(allPrecs)
print(table(number.selection))

number.selection
   1    2    3    4
5337   52    2    2
```

QC1 – Experiment-wide visualisation

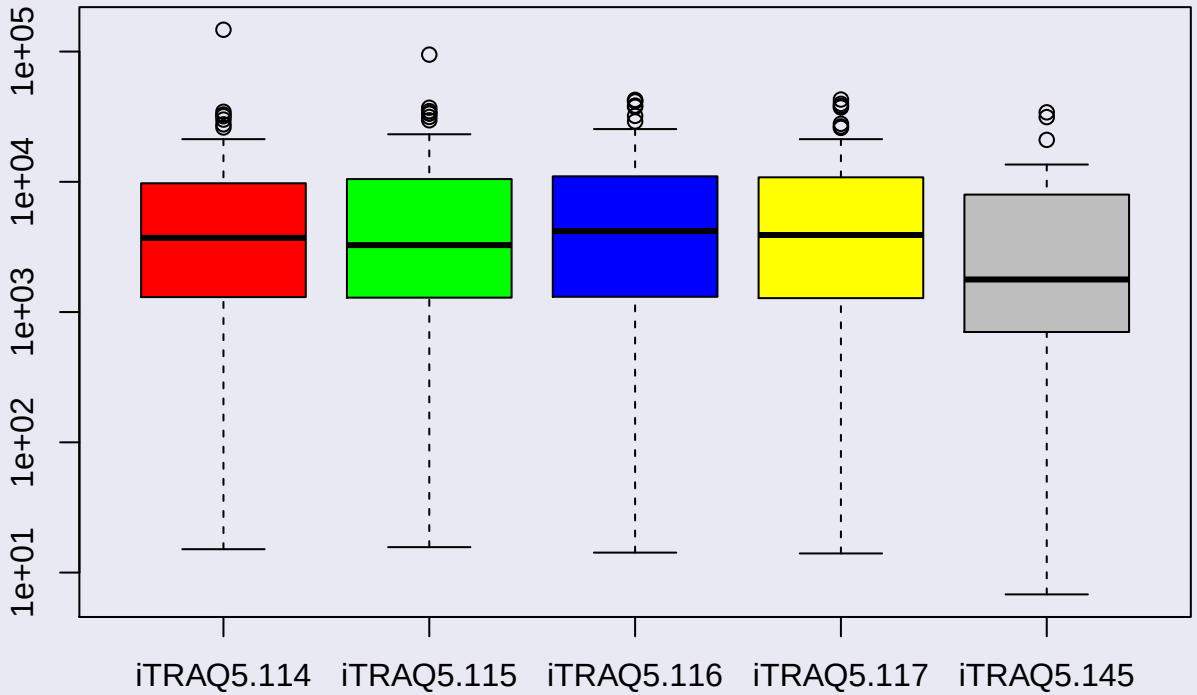


QC1 – Experiment-wide visualisation

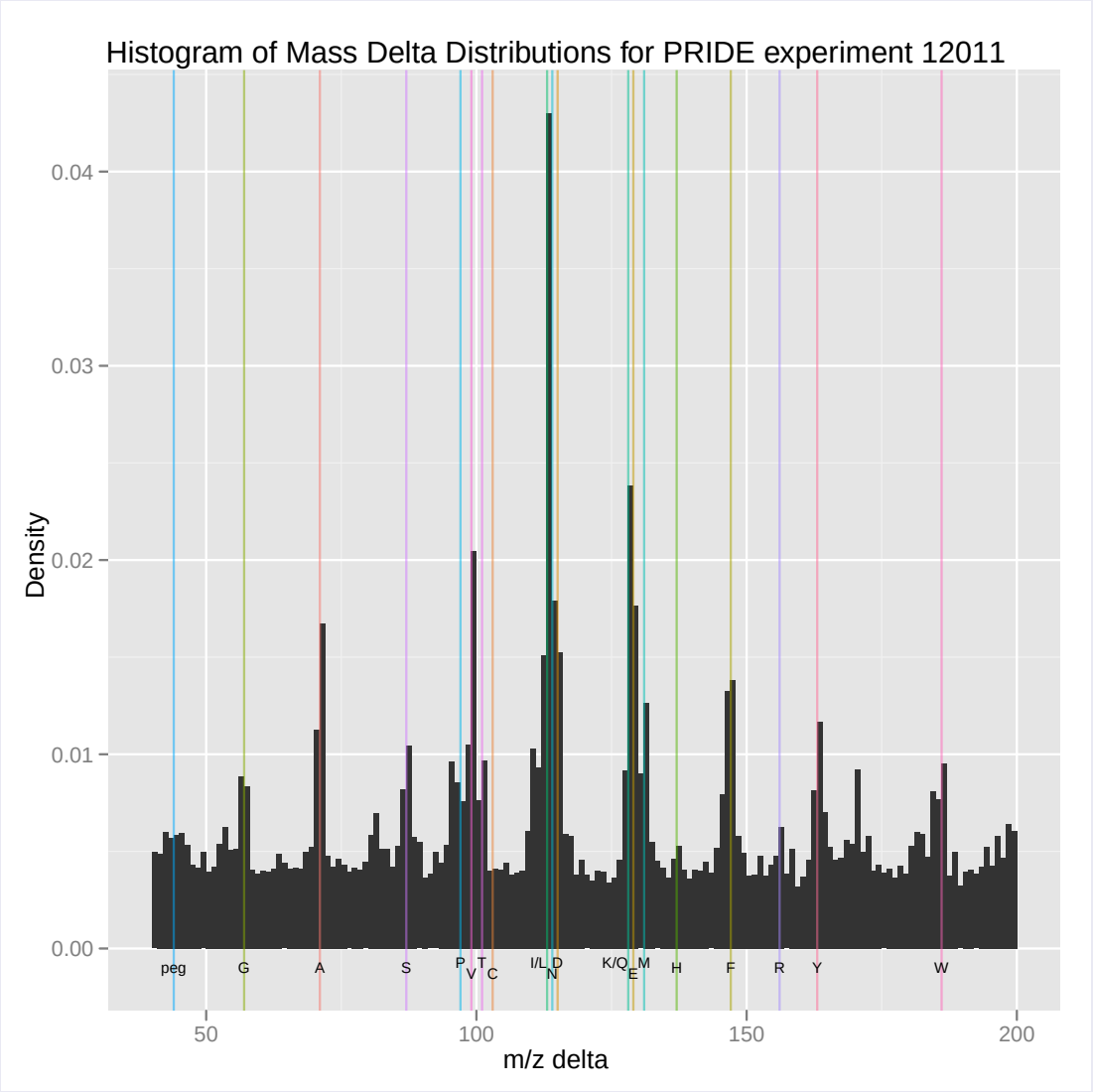


QC2 – Assessing incomplete dissociation

```
> foo <- quantify(itraqdata,"trap",iTRAQ5,verbose=FALSE)
> boxplot(exprs(foo),col=iTRAQ5@col,log="y")
```



QC3 – Spectra quality



Plan

`for (i in TODO)`

- On-disk random access of data (using proteowizard library) – `mzR` package under development with Bernd Fischer (EMBL) and Steffen Neuman (IPB HALLE, `xcms`).
- Some processing is embarrassingly easy to parallelise.
- Label-free quantitation.
- Easier integration of identification data.
- ...

More info, other packages

- `MSnbase` vignettes
- Proteomics sig mailing list – `https://stat.ethz.ch/mailman/listinfo/bioc-sig-sequencing`
- BiocViews – MassSpectrometry and Proteomics
- CRAN Task View – Chemometrics and Computational Physics

Acknowledgement

- Kathryn Lilley and CCP team.
- BBSRC Tools and Resources Development Fund Award.
- PRIME-XS FP7.

Thank you for you attention.