Pre-processing of metabolomics data

Benjamin Thiombiano

Frans van der Kloet

Overview

Introduction	• METABOLOMICS,(HP)LC, MS
Pre-processing steps	• What can go wrong and how to fix it?
Example	• Pre-processing metabolomics data

Metabolomics

Scientific study of chemical processes involving **metabolites**

Metabolites

- metabolic intermediates and products (usually <1500 Da in size)
  - Lipids, sugars, amino acids, ...
  - Exceptions: lipoproteins, albumin (can be detected with NMR)
- hormones and other signalling molecules
- secondary metabolites (in plant metabolomics; metabolites not related to growth, development and reproduction; e.g., antibiotics, pigments)

The **metabolome** represents the collection of all metabolites in a biological cell, tissue, organ or organism, which are the **end** products of cellular processes

Liquid chromatography (LC)

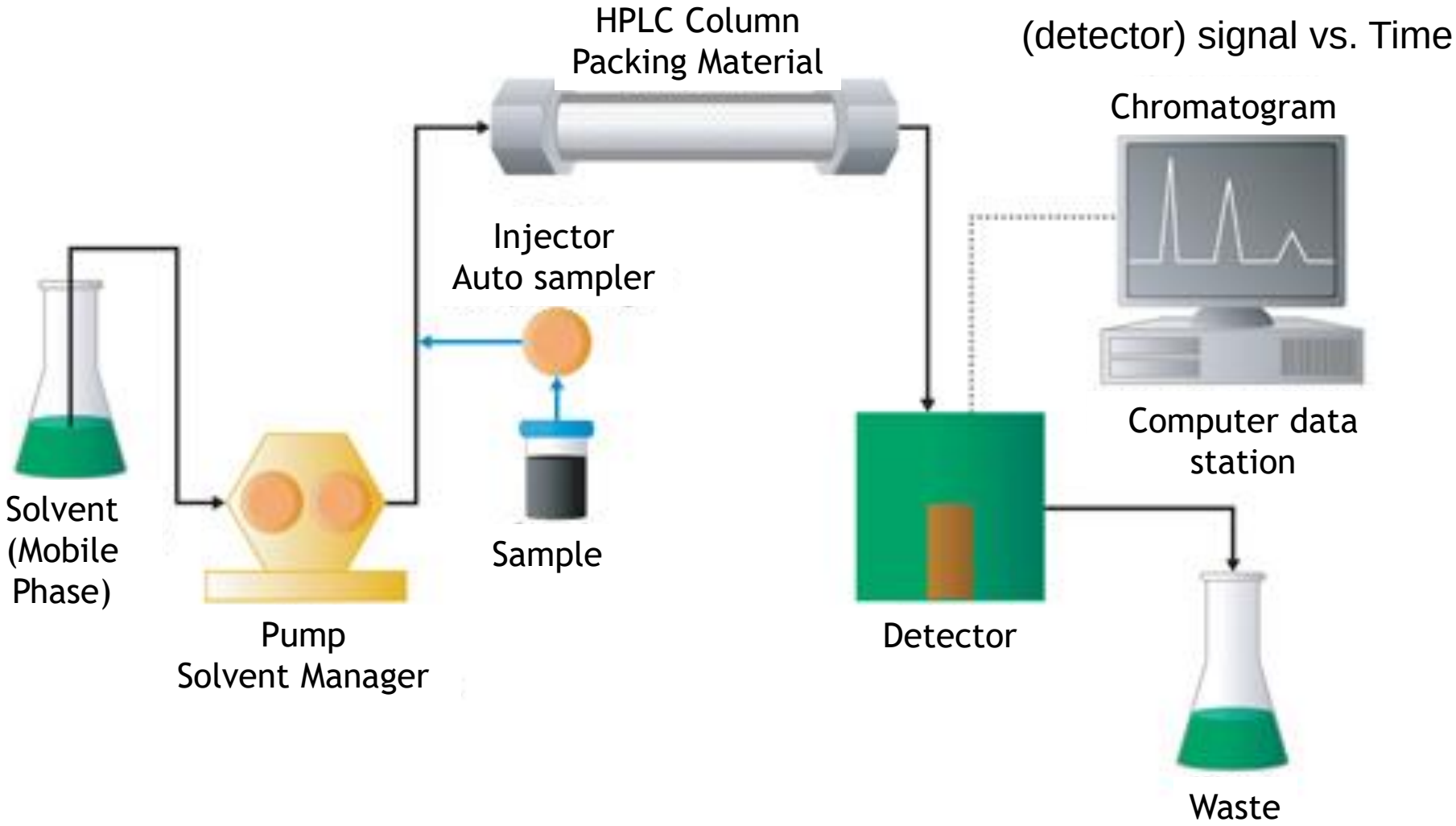
Analytical chromatographic technique for separating ions or molecules that are dissolved in a solvent.

Sample (solution) is in contact with a second solid or liquid phase

The different solutes will interact with the other phase to differing degrees due to differences in adsorption, ion-exchange, partitioning, or size.

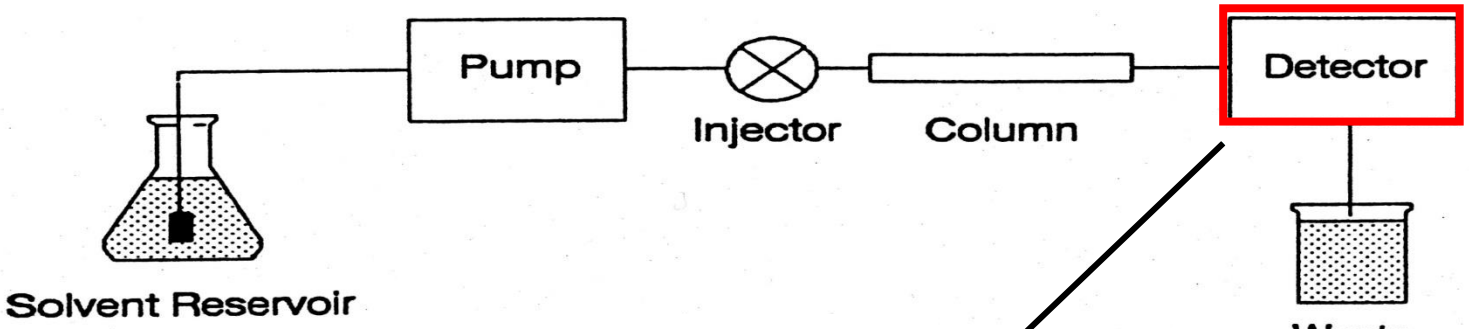
This interaction allows the mixture components to be separated

(HP)LC: How does it work?

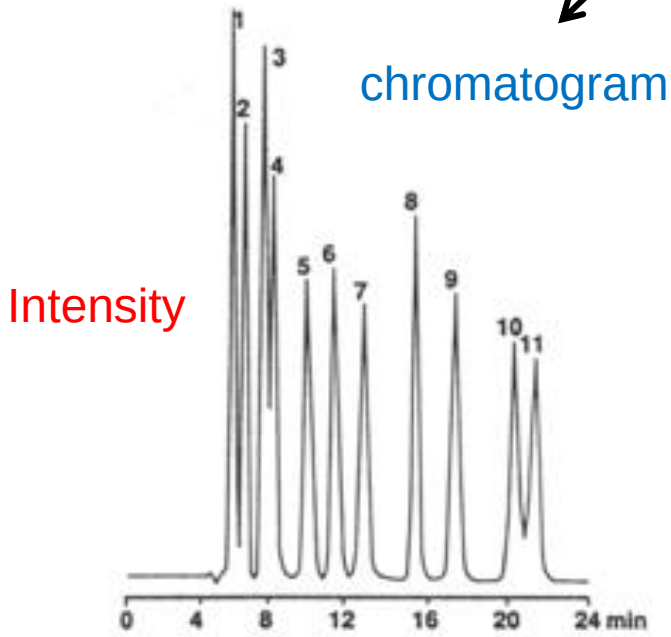


http://www.waters.com

Liquid Chromatograph: summary



e.g. Standard mixture of sugars and alcohols



1. Stachyose
2. Raffinose
3. Sucrose
4. Lactose
5. Glucose
6. Galactose
7. Fructose
8. Erythritol
9. Mannitol
10. Galactitol
11. Sorbitol

These peaks are identified. But how?

Mass Spectroscopy (MS)

1

MS is an analytical technique for the determination of the **composition** of a sample or molecule.

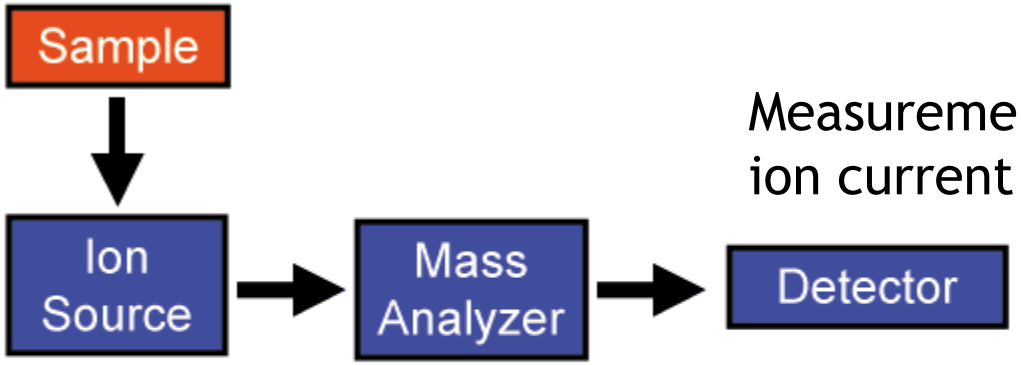
2

Used for elucidating the **chemical structures** of molecules, such as peptides and other chemical compounds.

3

The MS principle consists of ionizing chemical compounds to generate **charged molecules or molecule fragments** and measurement of their mass-to-charge ratios (**m/z**).

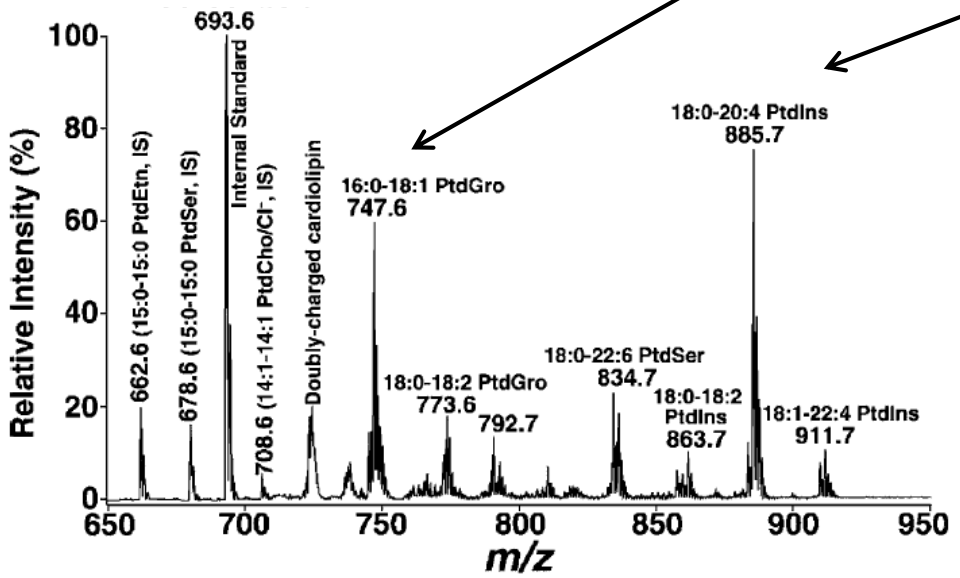
Mass Spectroscopy (MS)



Measurement of ion current

Masses are specific to chemical compounds

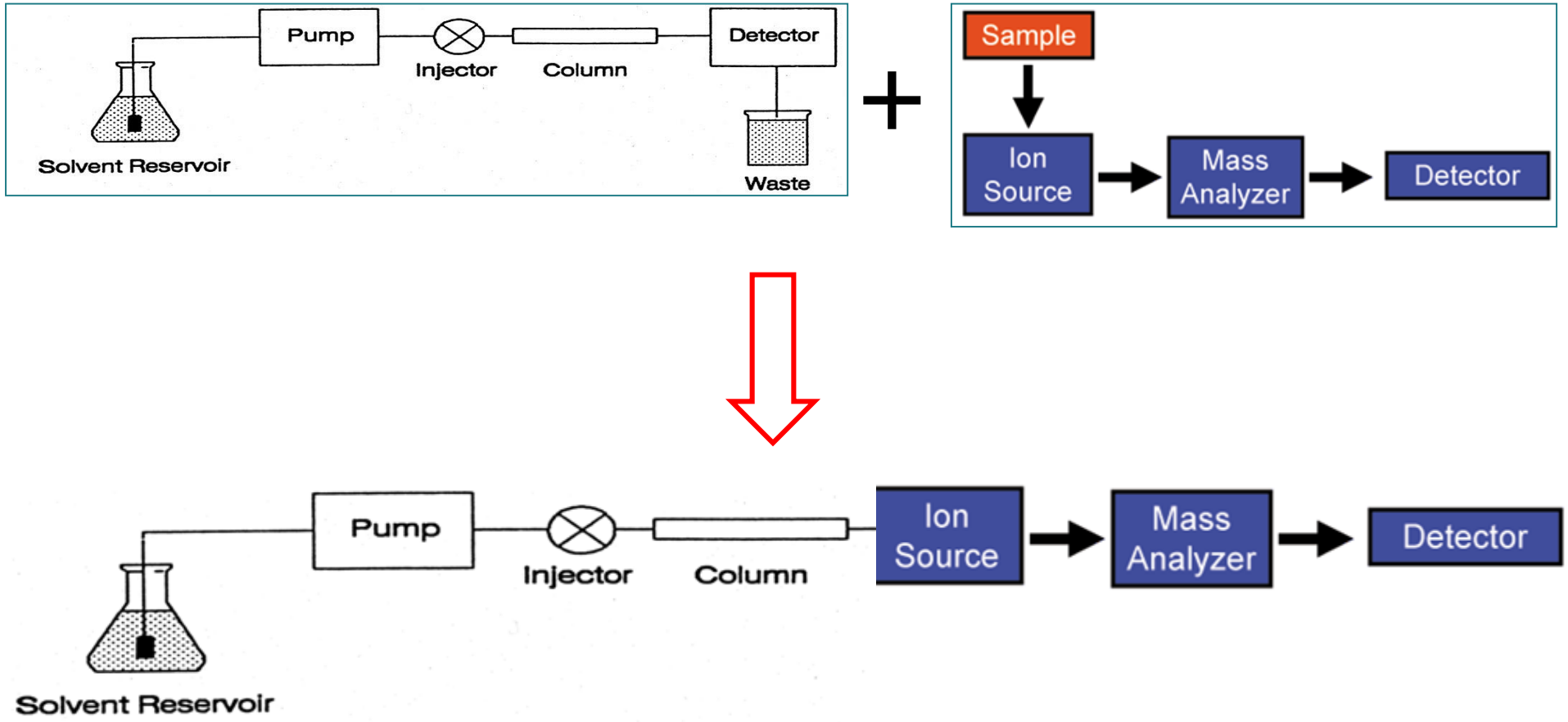
mass spectrum, mass scan (ion current)



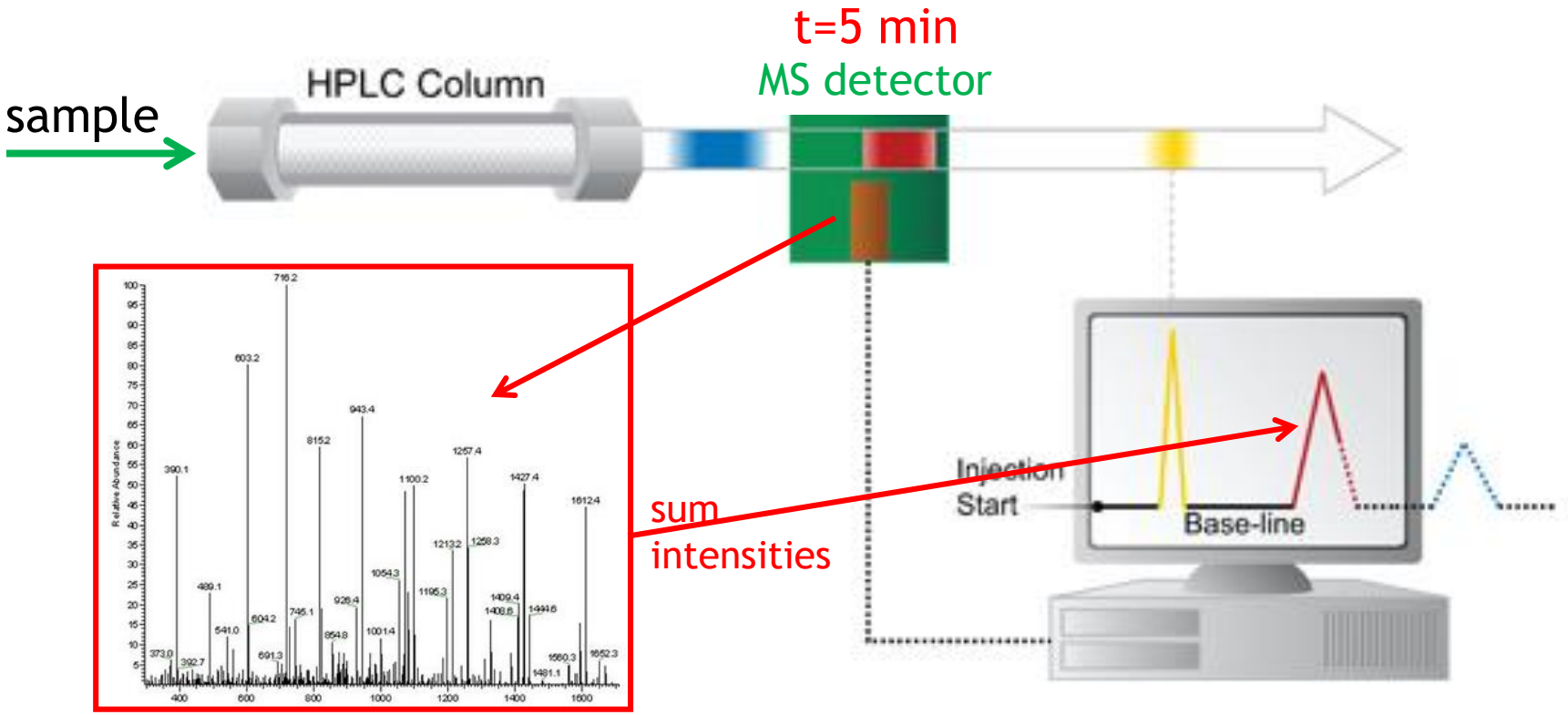
Metabolomics: Hyphenated MS (e.g. LC – MS)

Liquid chromatograph

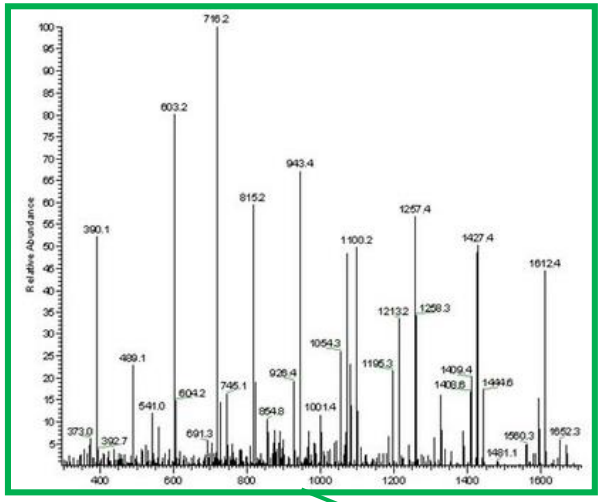
Mass - spectroscopy

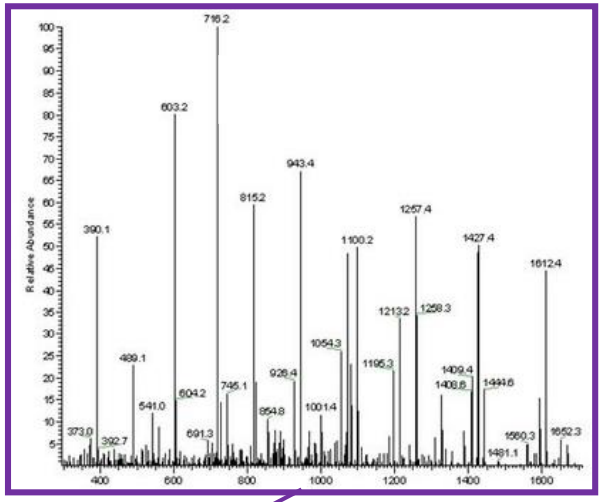


LC-MS:Measuring mass-spectra at regular time intervals



LC-MS data





sum intensities

sum intensities

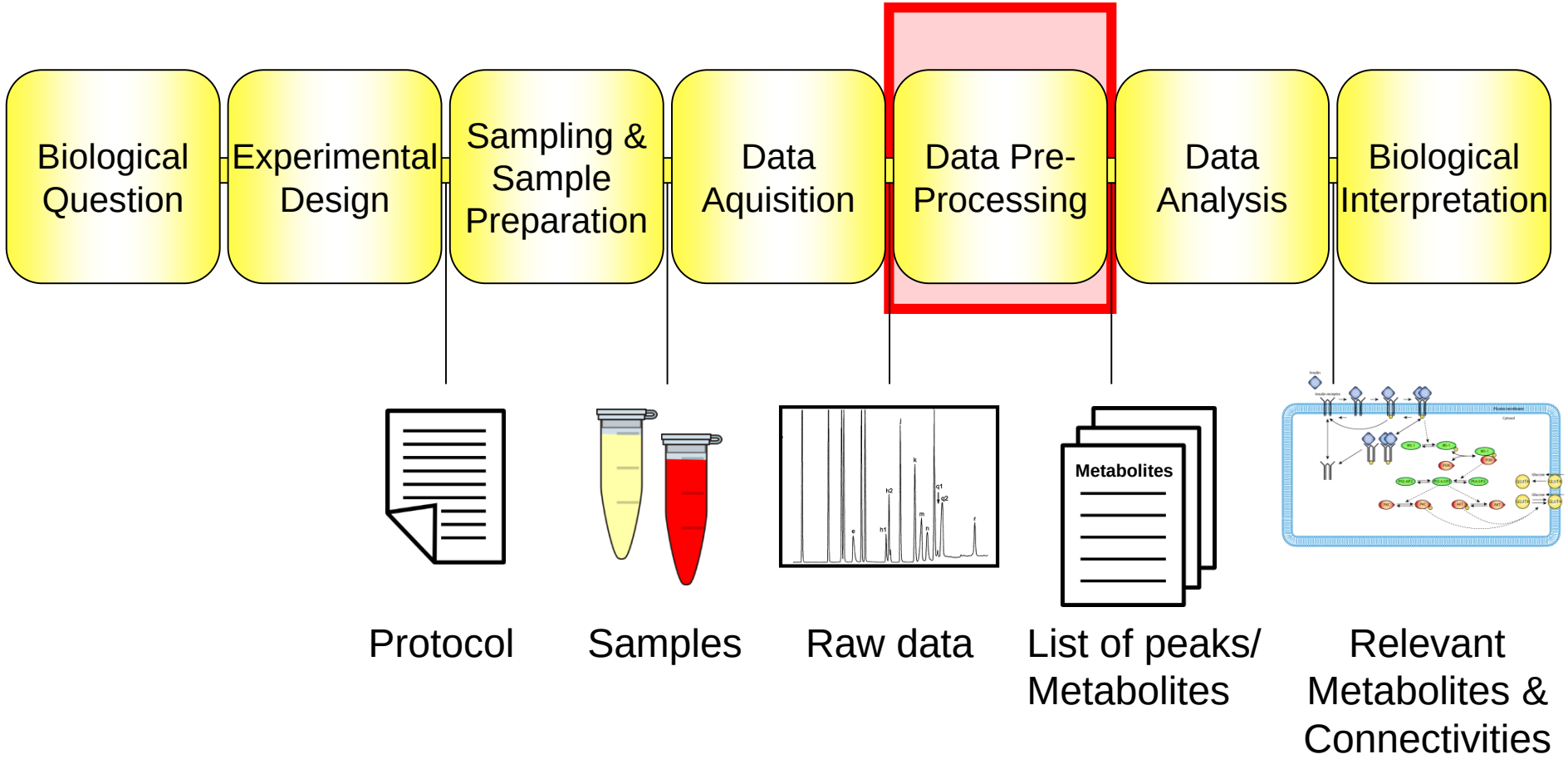
Total Ion Current (TIC)

t=120 s

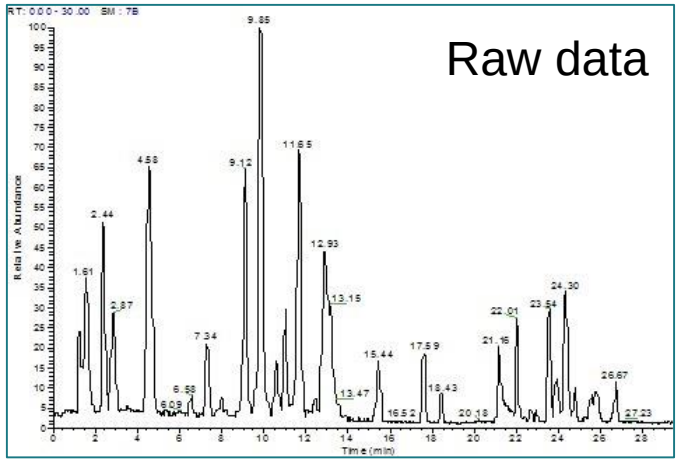
t=121 s

TIME

The metabolomics workflow



Data pre-processing



Data pre-processing

m/z	Rt (minutes)	Intensity	Start (minutes)	End (minutes)	Area	Metabolite
600.4	4	10000	3.5	4.4	50000	glucose
700.9	4.5	5000	3.6	4.6	10000	unknown
756.5	6	12056	5.6	6.4	34000	unknown
etc						

Peak table: list of peaks/metabolites

Correcting data from mass spectra

Data pre-processing

Need for correction (1)

Sample handling variation

- Preparation Inaccuracies
  - Weighing
  - Pipetting
- Gross errors
  - Forgotten to add (internal) standard
  - Wrong volumes added
  - Labelling errors

Chromatographic errors

- Carry-over and background effects
- Column aging
- Column changes
- Solvent changes

Need for correction (2)

Detector Limitations

- Low: at or just above LOQ
- High: detector saturation

Integration errors

- Wrong Peak
- Peak not in Window
- Inadequate peak integration method

Correction levels

01

Low level

- Mass correction
- Peak alignment

02

Sample level

- Integration
- Calibration lines
- Internal standards

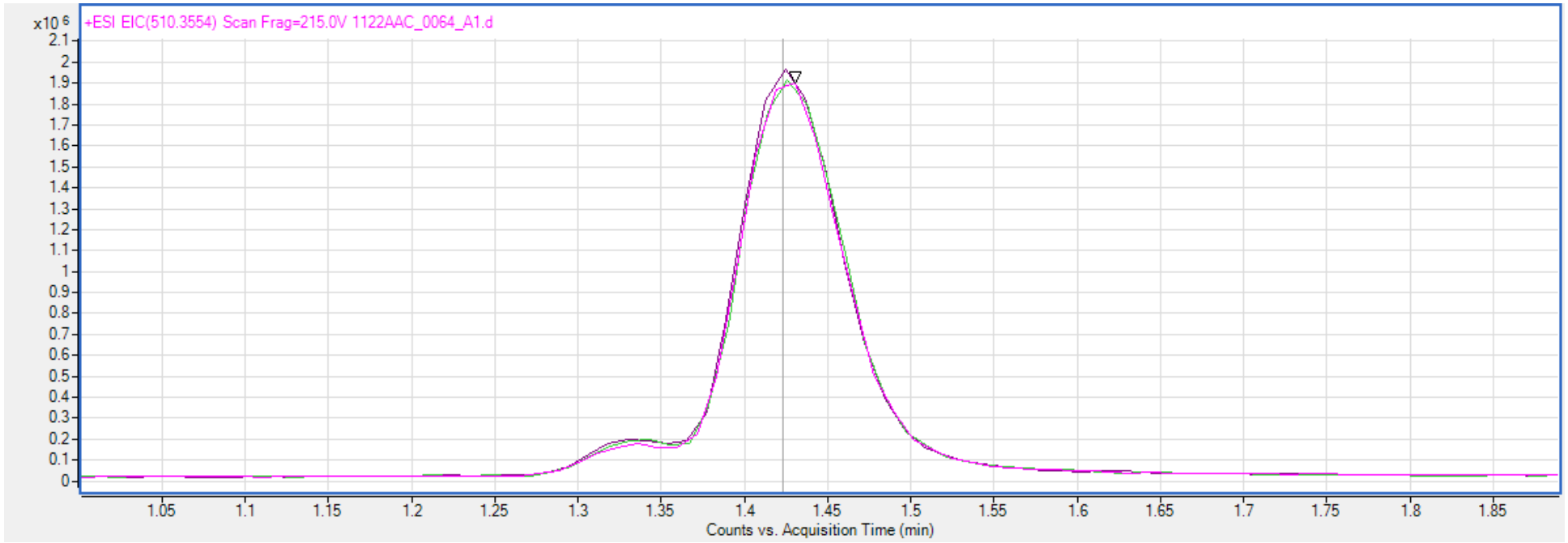
03

Batch level

- Reference samples
- QC correction

Integrate the chromatogram

Extracted Ion Chromatogram of 510.3554 Da

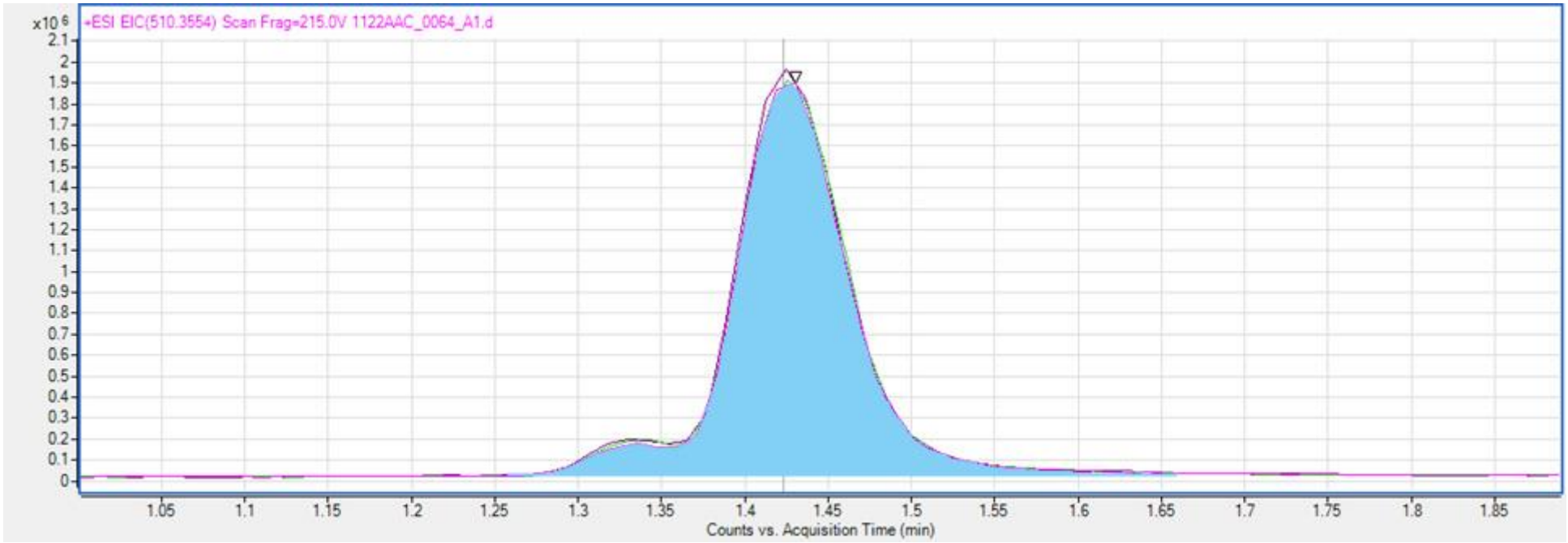


Retention time (scans) ->

Notation = **mz@rt**, 510.3554@1.425 = a single feature!

Integrate = sum area under curve

Extracted Ion Chromatogram of 510.3554 Da

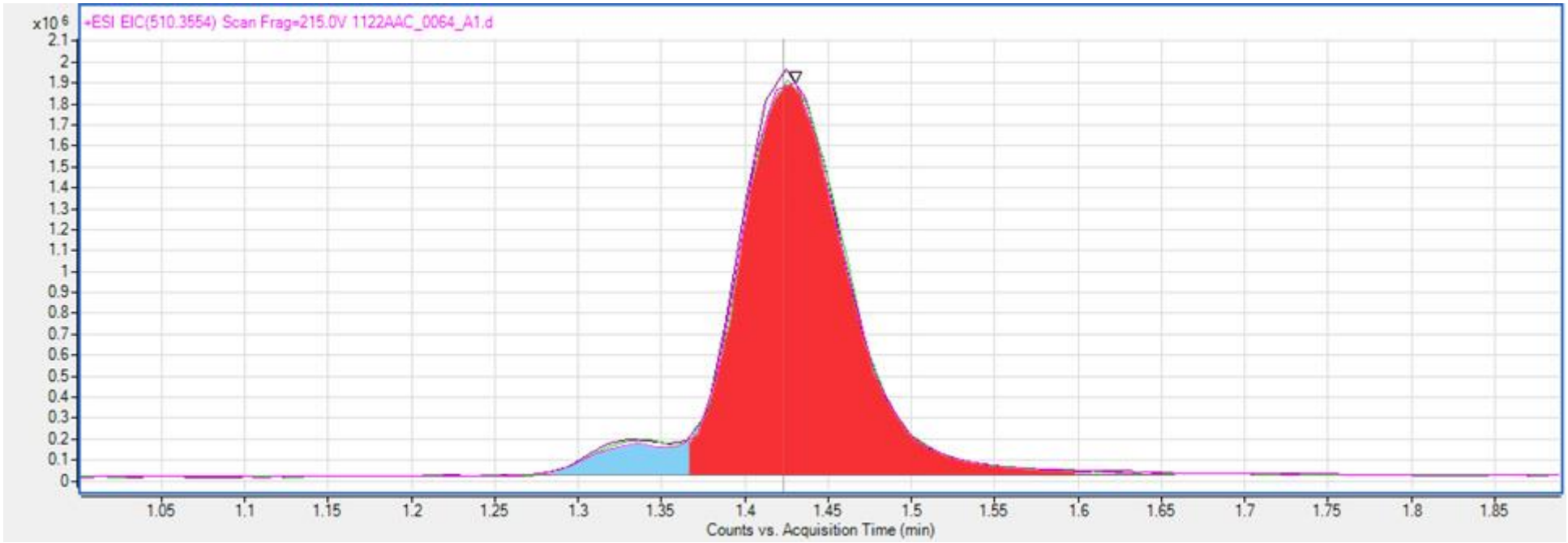


Retention time (scans) ->

Notation = **mz@rt**, 510.3554@1.425 = a single feature!

Or of course two peaks

Extracted Ion Chromatogram of 510.3554 Da

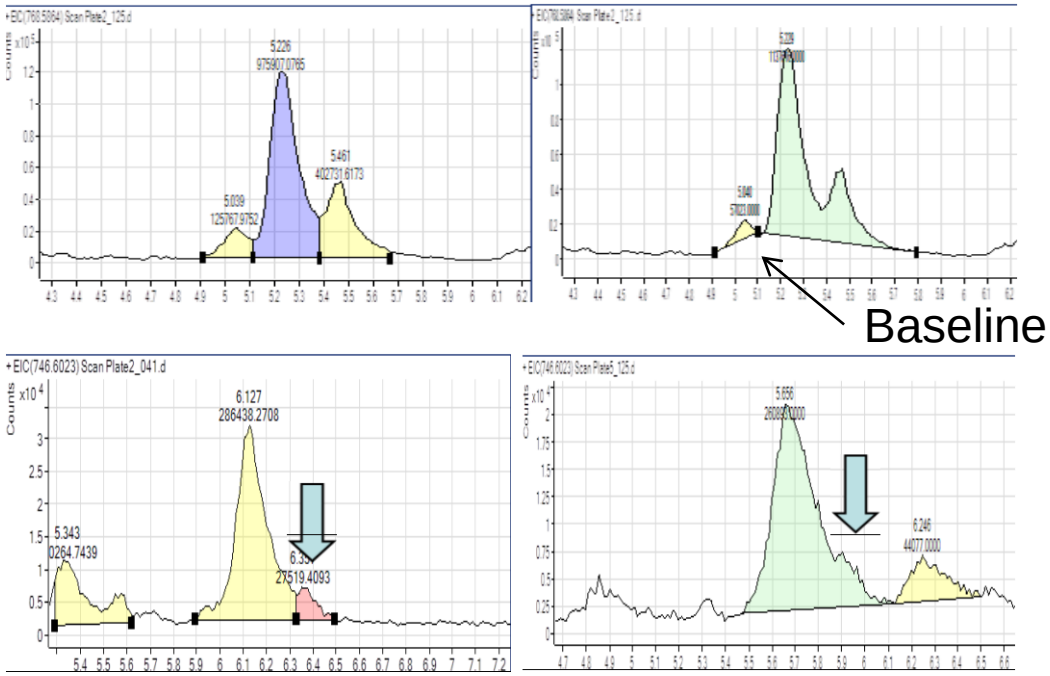


Retention time (scans) ->

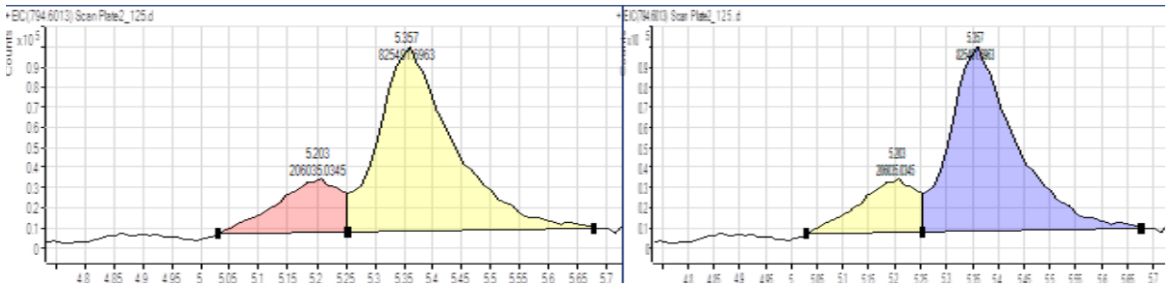
Notation = **mz@rt**, 510.3554@1.425 = a single feature!

Integration errors

peaks are separated or combined



Baseline levels?



Wrong peak is integrated

Is it a big problem?

Targeted analysis

- Over time the integration parameters for all peaks are optimized/fine-tuned.
  - Data is manually curated (~1-2 days per batch (150 samples))
- Software gets updates (feedback from users to vendor)

Untargeted analysis

- Yes, there is no luxury of knowing what to look for so no integration parameters to optimize.

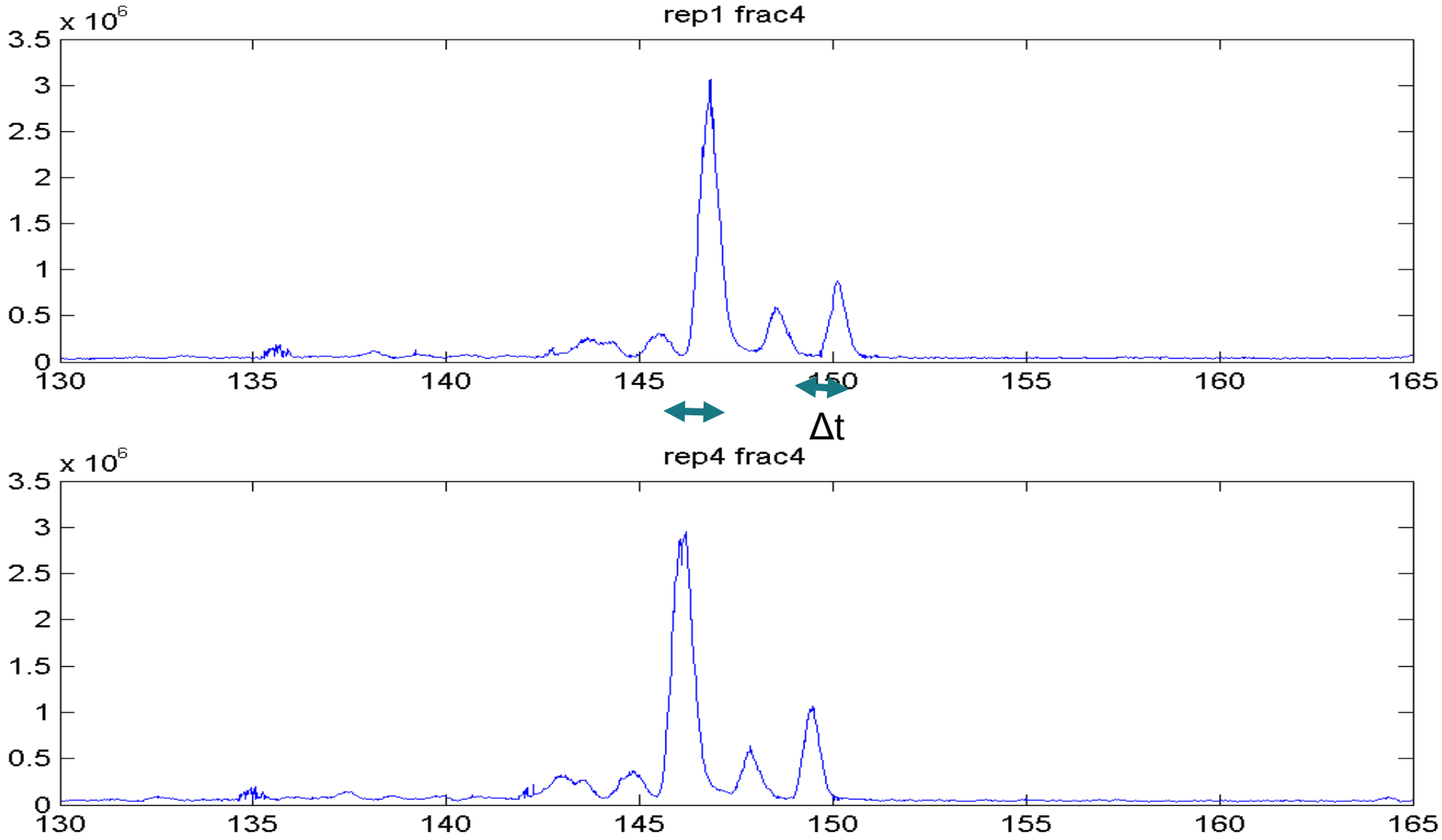
We have peak areas, what about retention time?

Match peaks sample A with peaks sample B, C, ... (all samples)

Problem retention time shifts due to

- small variations
  - in temperature,
  - atmospheric pressure,
  - pH, time and composition of the samples
- column degradation

Retention time shifts



Retention time shifts

Automated tools (small local time shifts):

Warping / alignments methods like *Correlation Optimized Warping* (**COW**), *Dynamic Time Warping* (**DTW**), *Parametric Time Warping* (**PTW**) are available.

Global working procedure:

-Split target spectrum up in multiple parts.
-Shrink and stretch parts in such a way that similarity with reference spectrum is optimized.

Caveats when using hyphenated-MS

Each metabolite has its own response factor

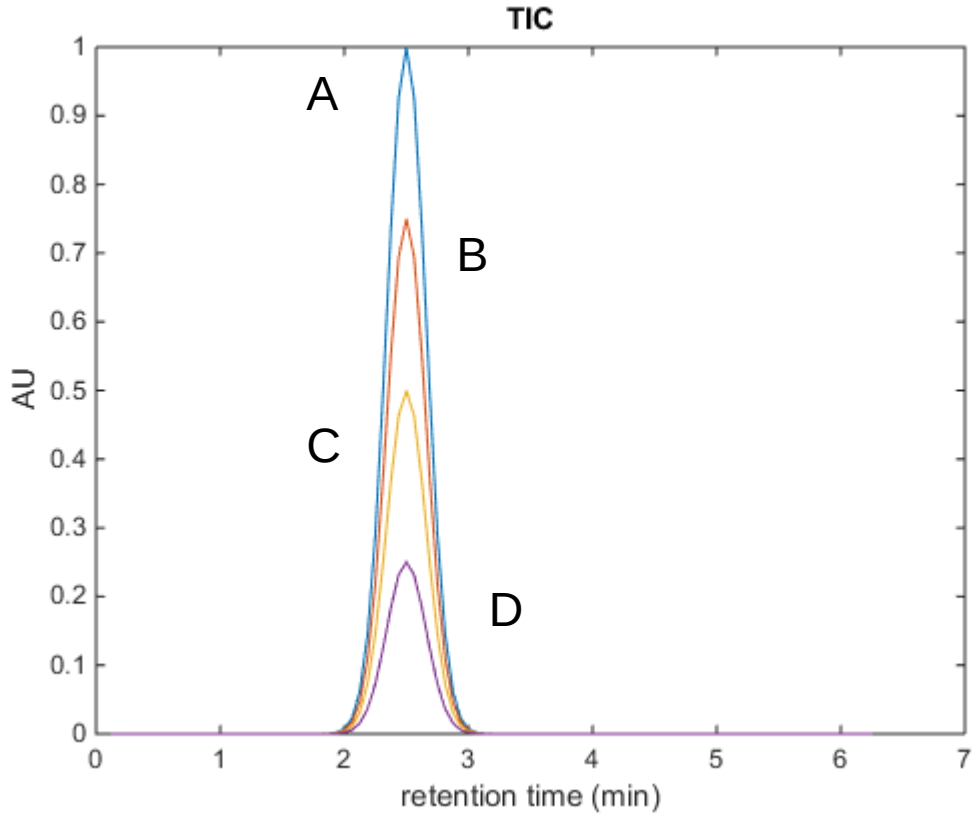
- i.e. the signal depends on the number of molecules but also on the type of molecule. This depends on factors like solubility, ionizability, fragmentation, mass discrimination at detector level.

The response factor is matrix dependent

- Different samples have different matrices.
  - Two samples with the same concentration can have different responses.
- The matrix is not constant over time.
  - Because of the chromatography step the matrix of the sample changes continuously and another source of variation is introduced.

Internal standards needed !!!

Own response factor

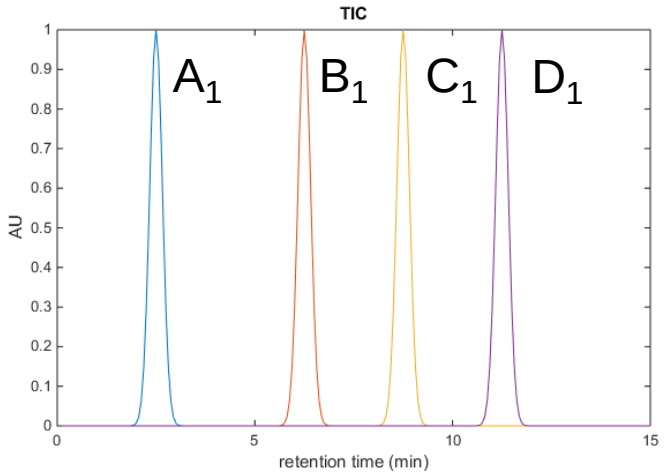


- 1 sample
- 4 peaks (A,B,C and D)
- Different area under the curves

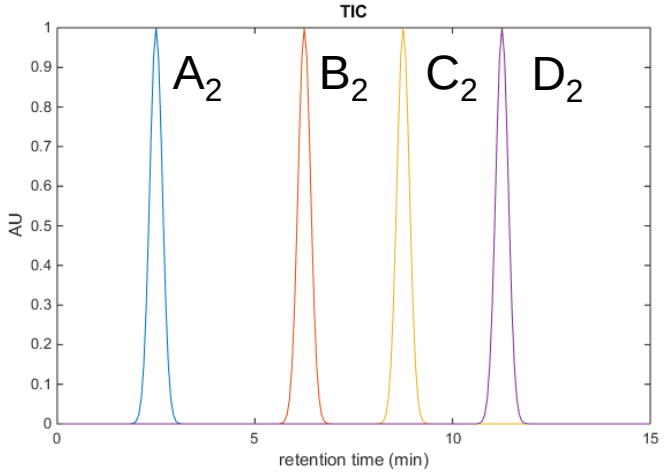
- E.g. [D]>[A] & [C]>[D]

Different matrix effects

Sample 1



Sample 2



- 2 samples
- 4 peaks per sample
- same area under curve

- $[A]_1$!= $[A]_2$
- $[B]_1$!= $[B]_2$
- $[C]_1$!= $[C]_2$
- $[D]_1$!= $[D]_2$

Internal Standards

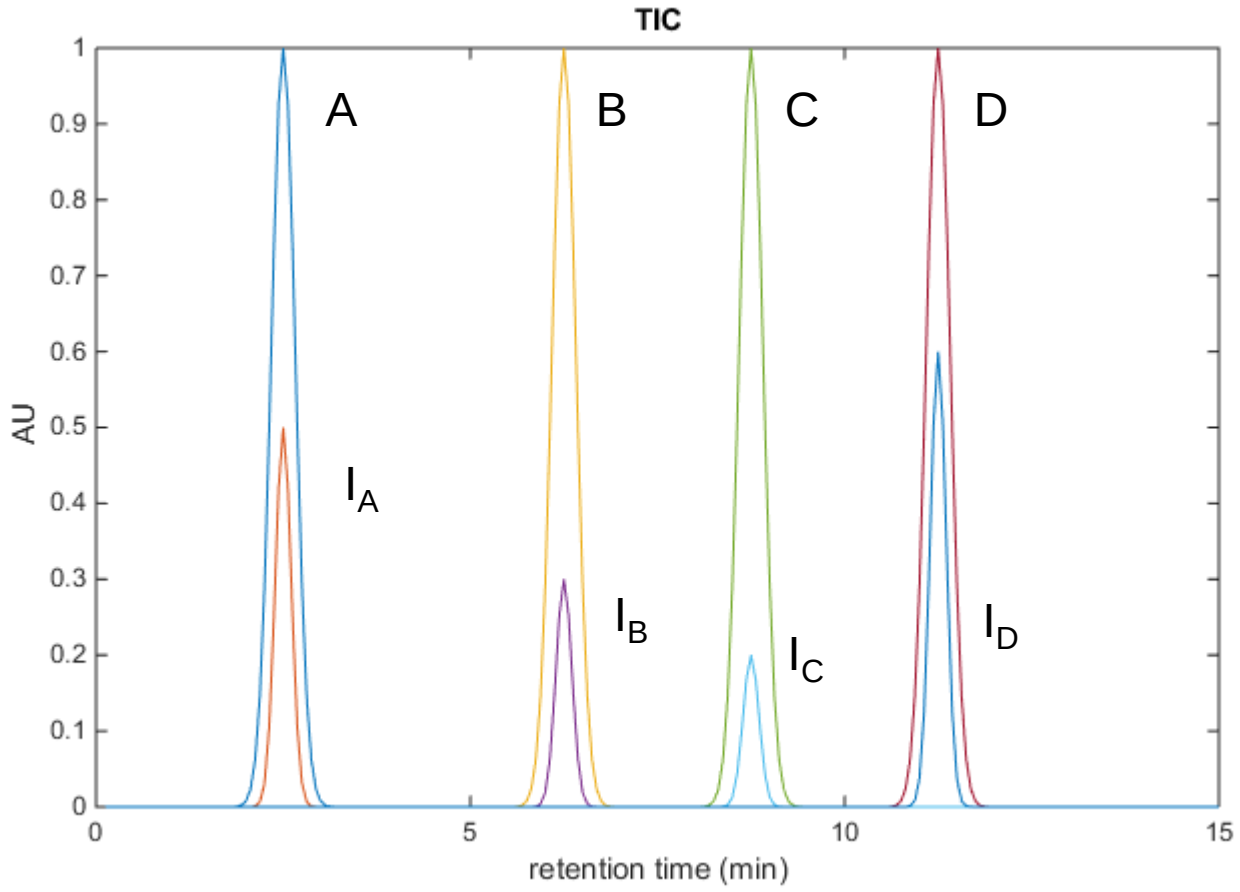
Function

- Monitor and/or correct variation introduced between sampling, sample preparation and sample acquisition.
- Variation sources sample preparation: filtration, pipetting, centrifugation, derivatization, etc. etc.

How?

- Near identical chemical behavior as the analyte
- In mass spectrometry:
  - provides its own ions
  - usually a stable-isotope labeled (C^{13} or N^{15}) compound as standard, or
    - isomers
    - homologues
    - structural analogs
- Sources of variation should affect both sample and IS the same

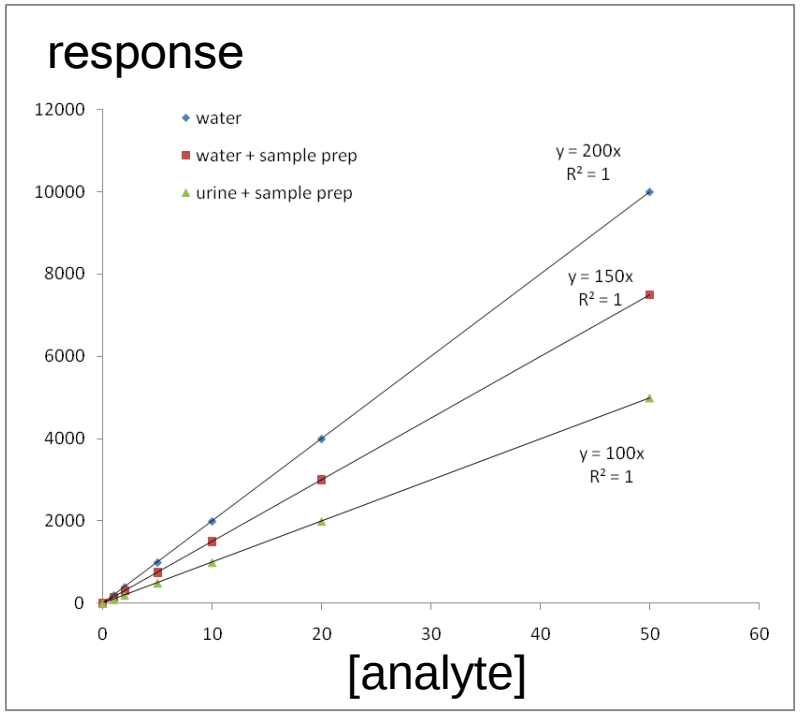
Example



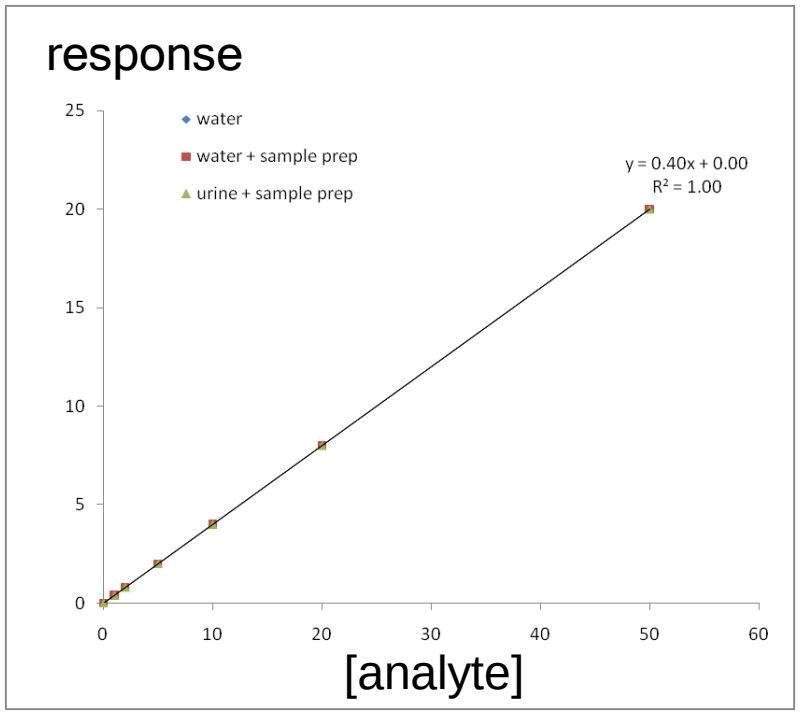
- 1 sample
- 4 peaks (A,B,C and D)
- 4 Internal standards
- 4 Ratios !!

Effect IS

Uncorrected

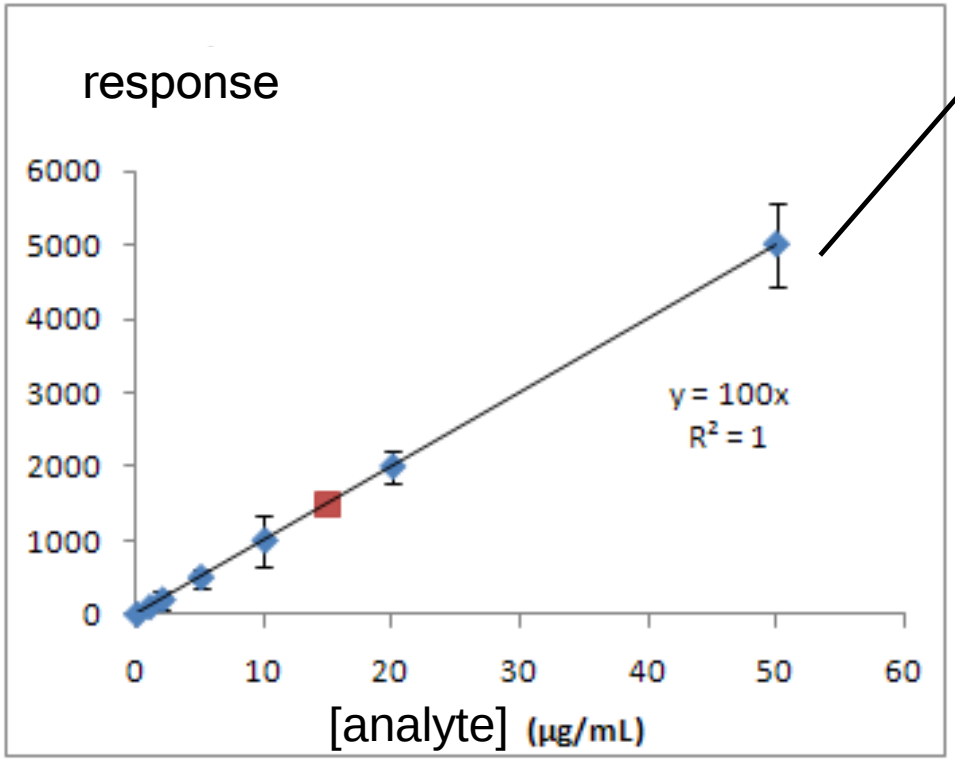


Corrected with IS



Repeatability

Measure replicates



error bar of standard deviation measure in triplicate!

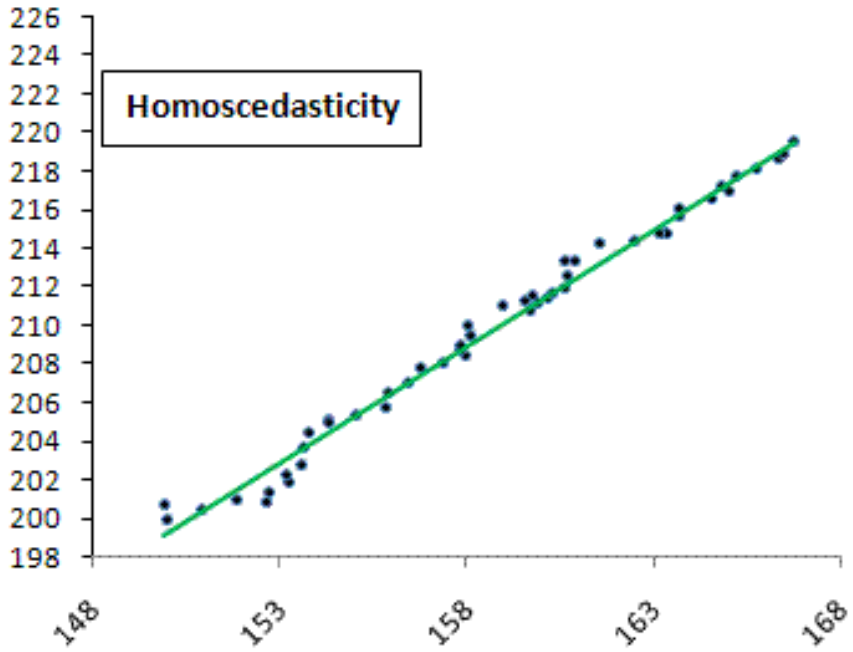
$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

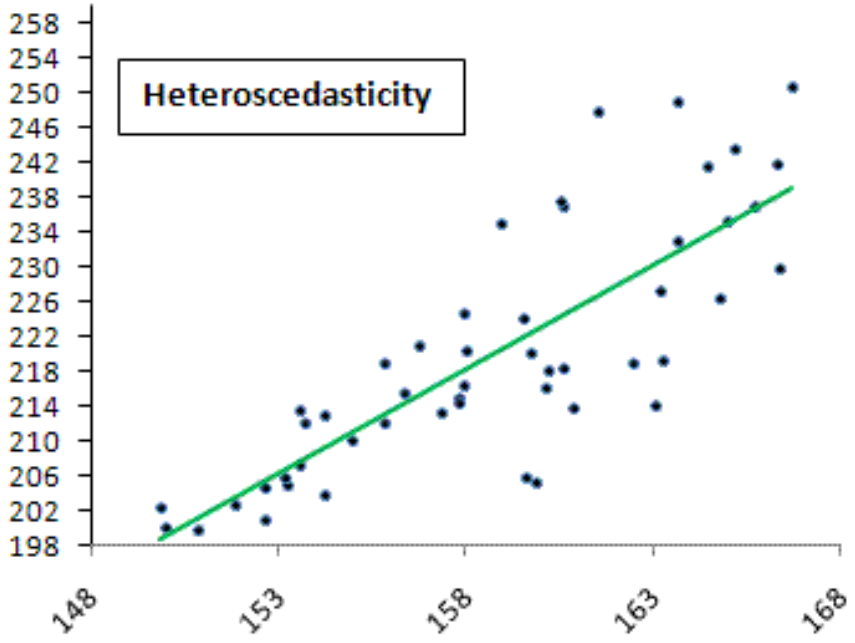
$$W = \frac{1}{s^2}$$

Correct for heteroscedasticity

$$(X^T X)\hat{\beta} = X^T y$$

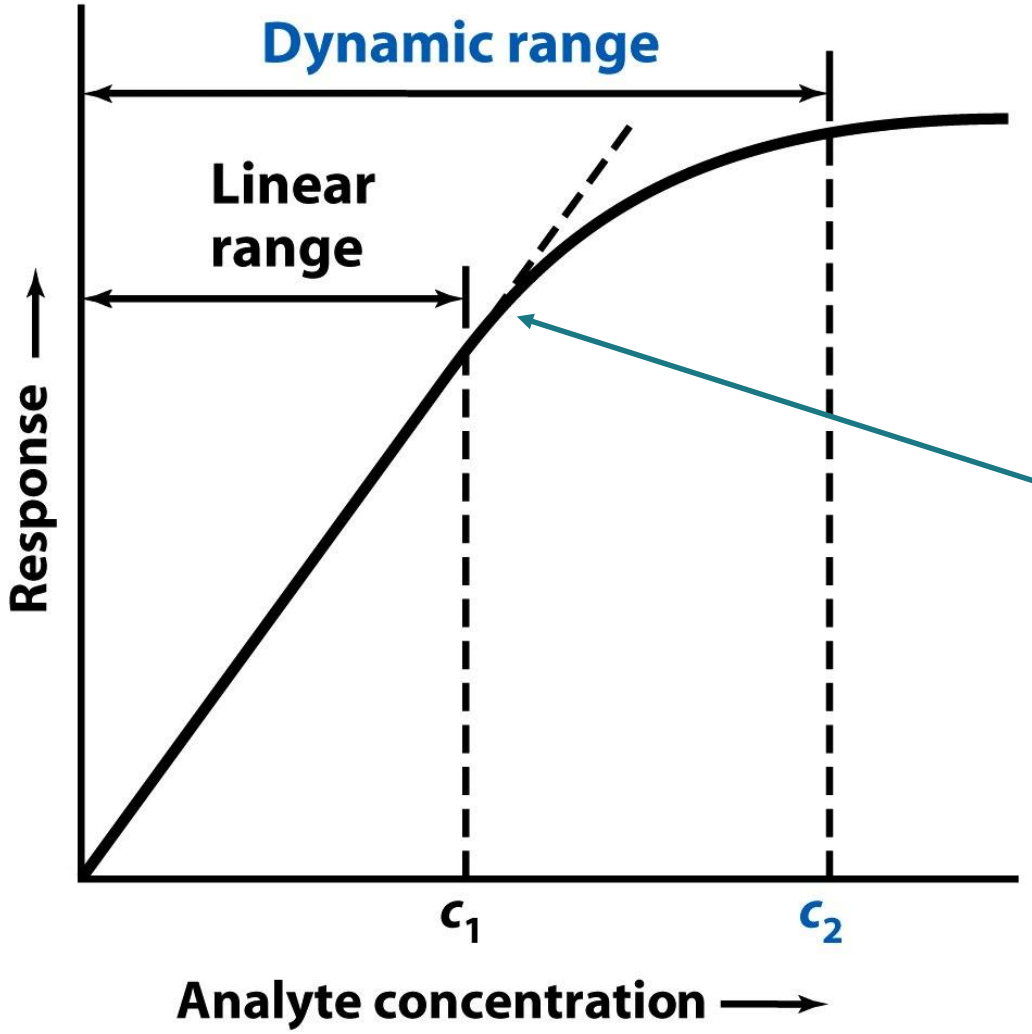
$$(X^T W X)\hat{\beta} = X^T W y$$





http://www.applyingdata.com/

Dynamic range vs. linear range



Saturation of detector
Extrapolation not allowed
Beyond this point

Figure 4-12
Quantitative Chemical Analysis, Seventh Edition

Correction levels

01

Low level

- Mass correction
- Peak alignment

02

Sample level

- Integration
- Calibration lines
- Internal standards

03

Batch level

- Reference samples
- QC correction

Extra normalization/correction needed still

IS **cannot** correct for dilution effects of the biosample e.g. urine

For many compounds no internal standard. Use 1 internal standard per class of compounds.

IS **cannot** correct for things like column degradation, different machinery etc.

Another type of normalization (over samples) needed:

- stable compound: concentration is constant for all study samples
- Use average samples as reference samples => **QC samples**
- hypothesis: reference samples should behave the same over time and ratio compound/IS should remain constant

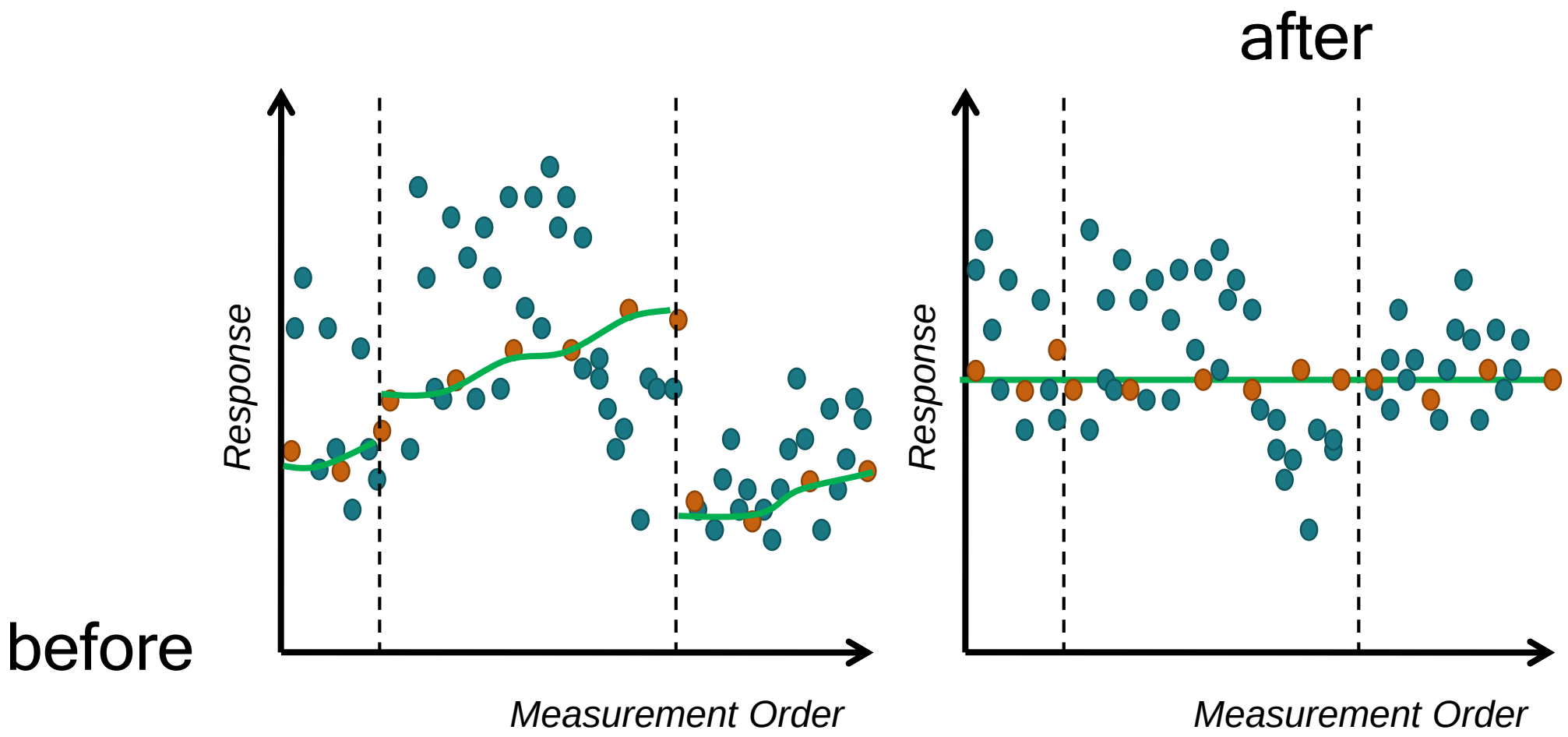
QCs and replicates

Sample acquisition variation (experimental drift) => QCs

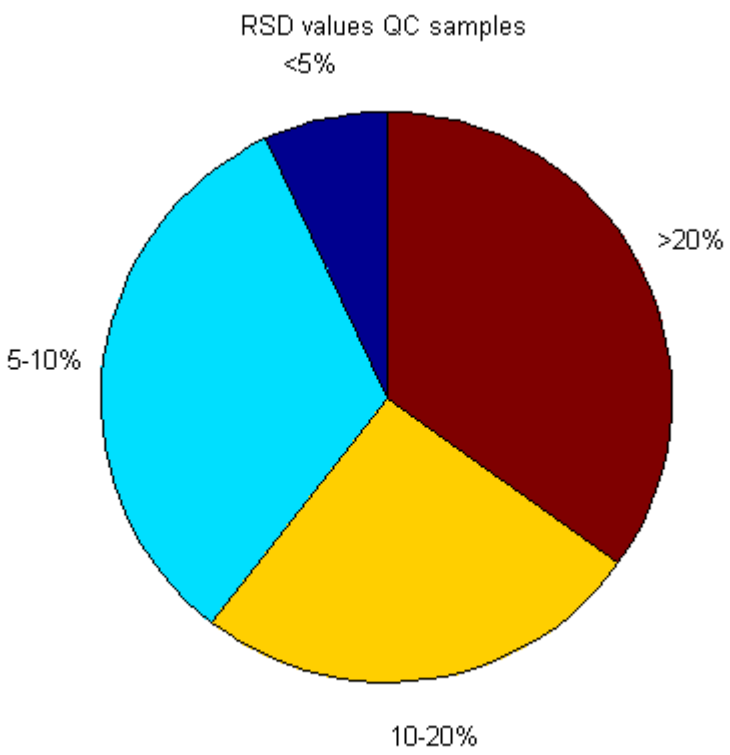
Sample preparation & acquisition variation => replicate samples

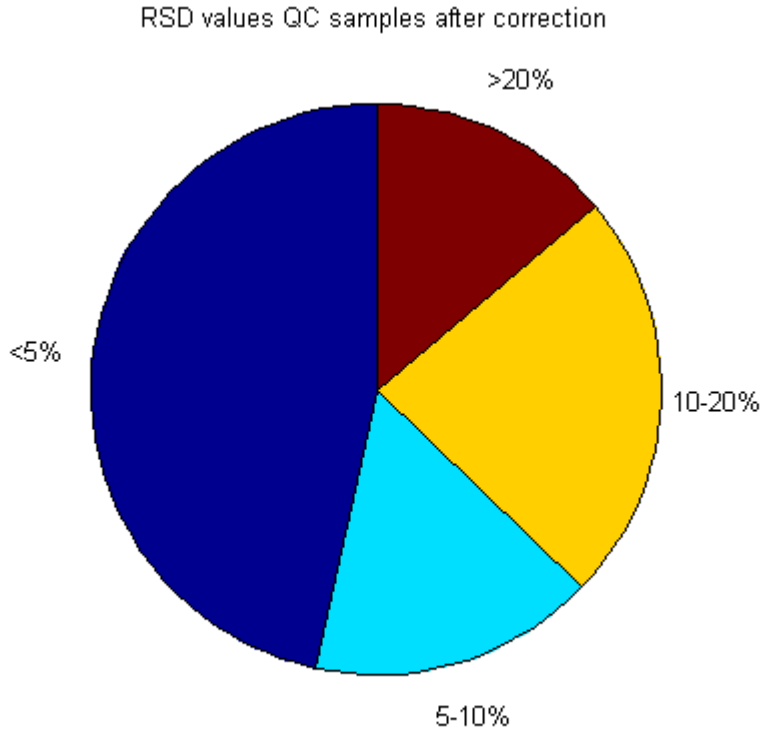
- injection replicates
- sample preparation replicates

QC correction



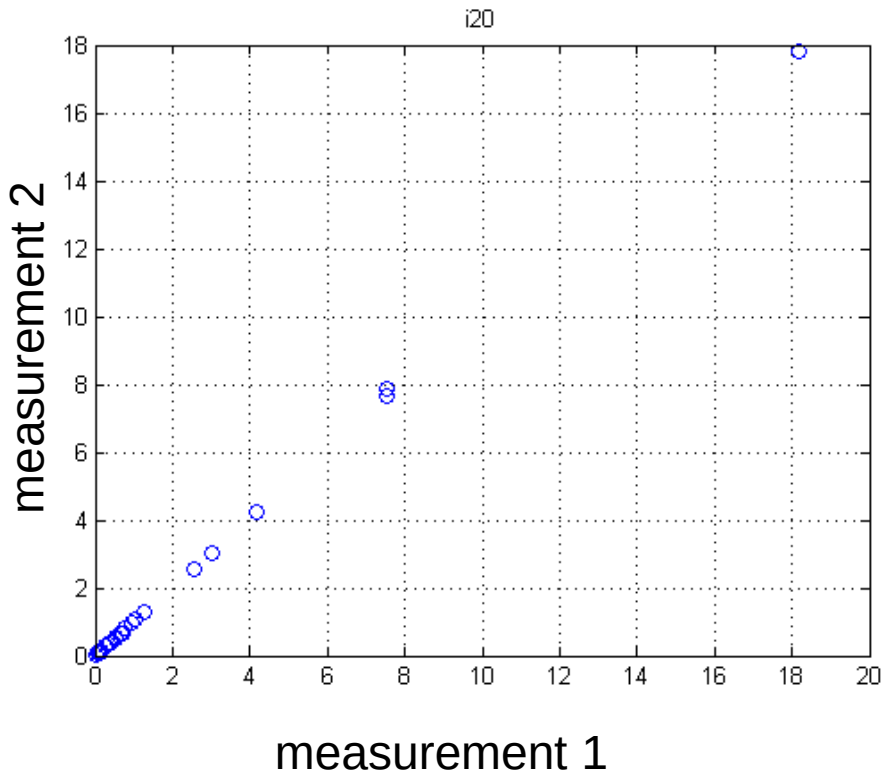
Quality Check (RSD QC samples)

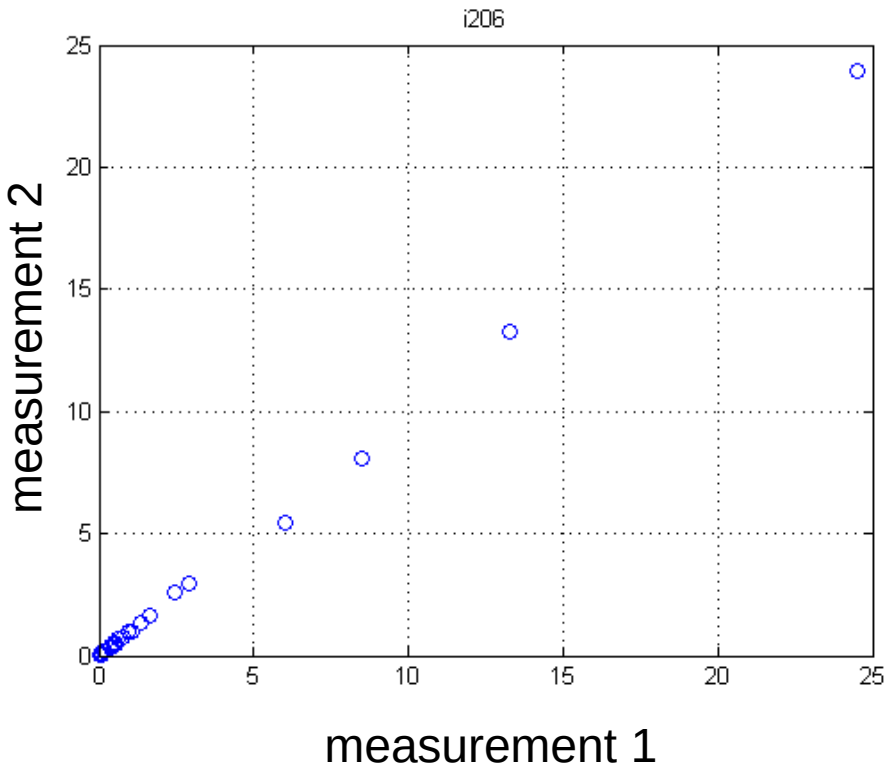




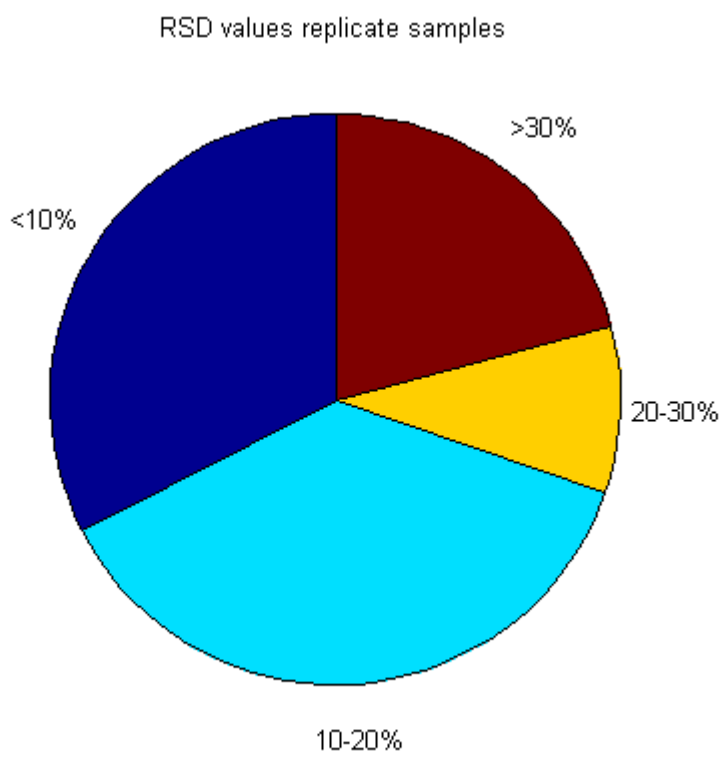
raw → corrected

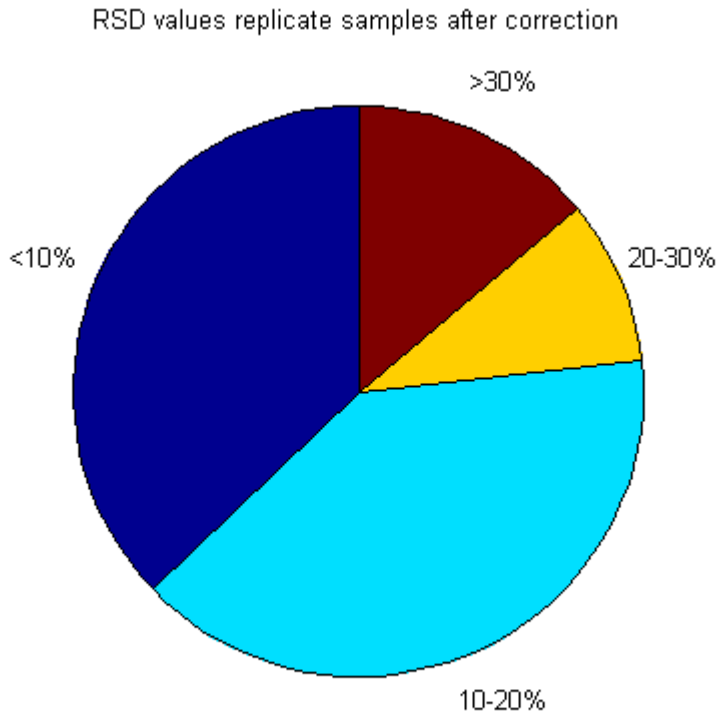
Replicates measurements



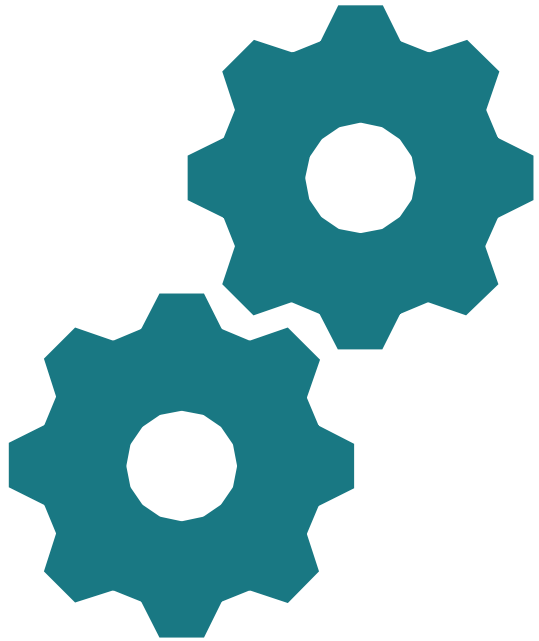


Quality Check (replicates)





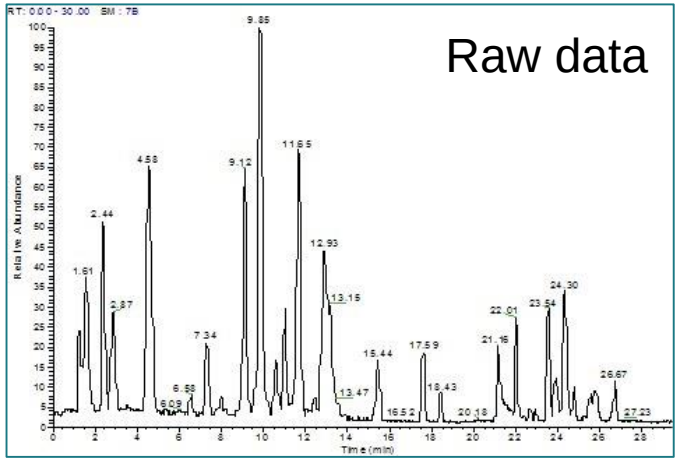
raw → corrected



Summary: standard operating procedure for pre-processing MS data

1. Integration (semi) automatic integration
2. Corrections:
    a. Blank correction
    b. Internal standard corrections
    c. Intra and inter batch (QC) correction
3. Quality control
    a. Create report
    b. Detect, inspect and correct deviating samples. When necessary, repeat steps 1-3b until desired quality
    c. Remove metabolites measured with insufficient quality from the data and export.
4. When possible and required: regression and quantification

So we can finally fill the peak table



Data pre-processing

m/z	Rt (minutes)	Intensity	Start (minutes)	End (minutes)	Area	Metabolite
600.4	4	10000	3.5	4.4	50000	glucose
700.9	4.5	5000	3.6	4.6	10000	unknown
756.5	6	12056	5.6	6.4	34000	unknown
etc						

Peak table: list of peaks/metabolites

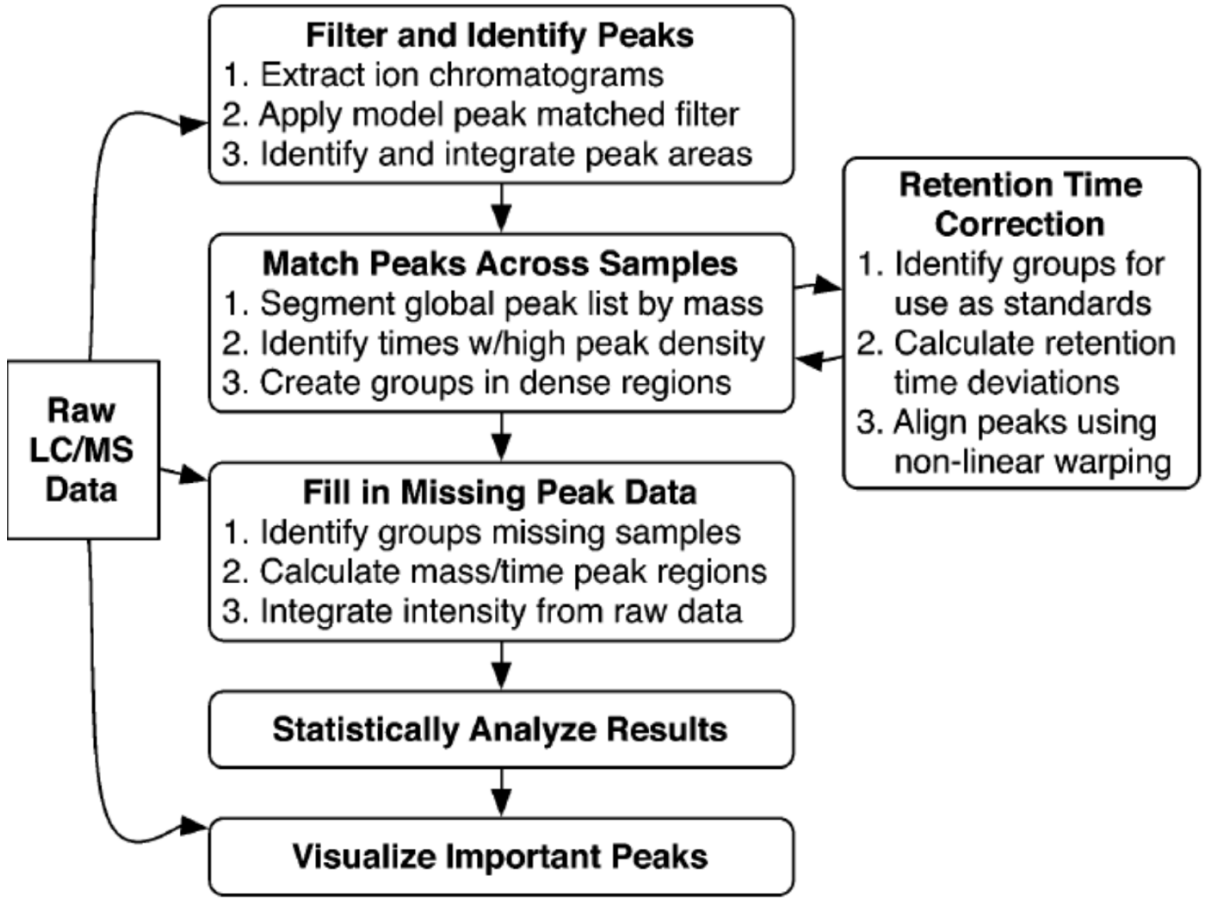
Example, What we want to do

Evaluate the effect of microbes on striga infection

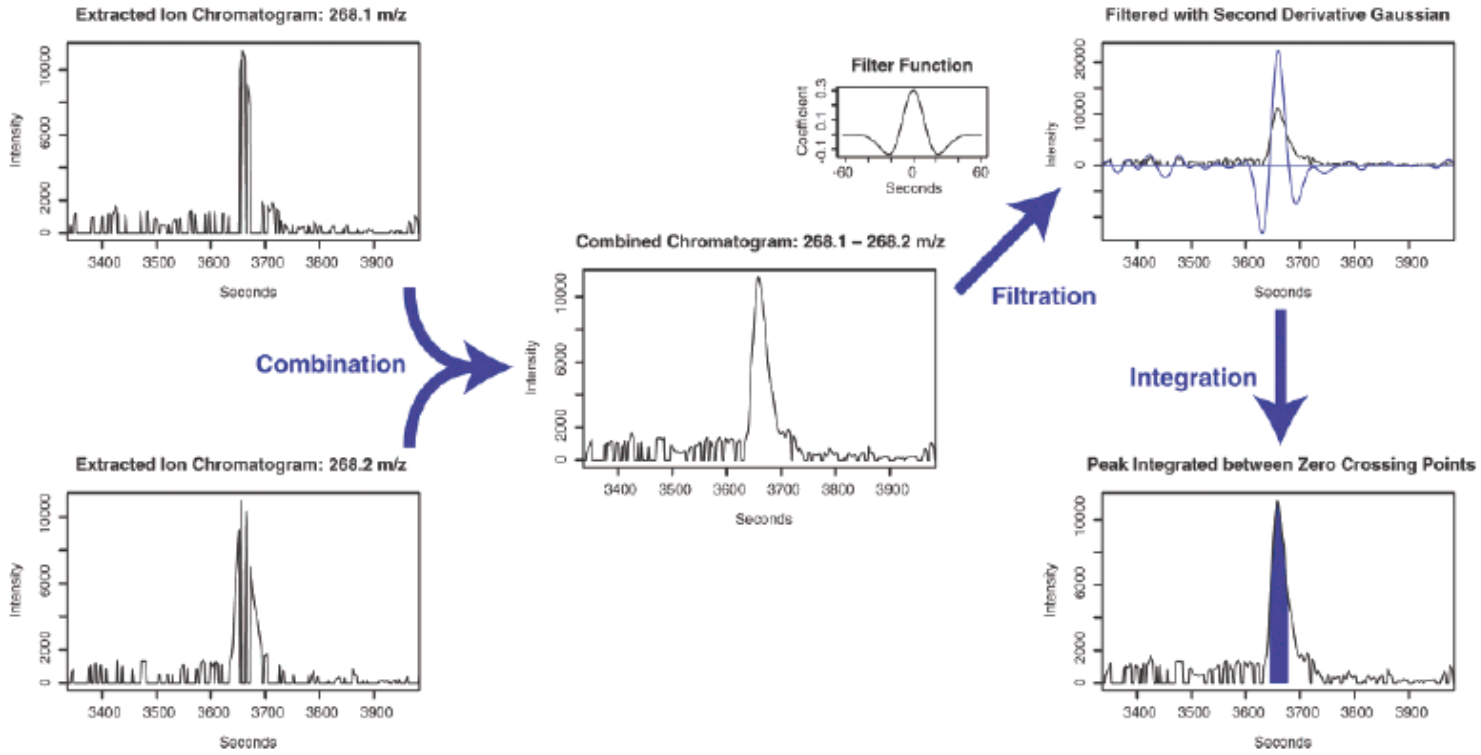
Are physiological modifications occurring in the plant?

The tool was to use metabolomics analysis of roots

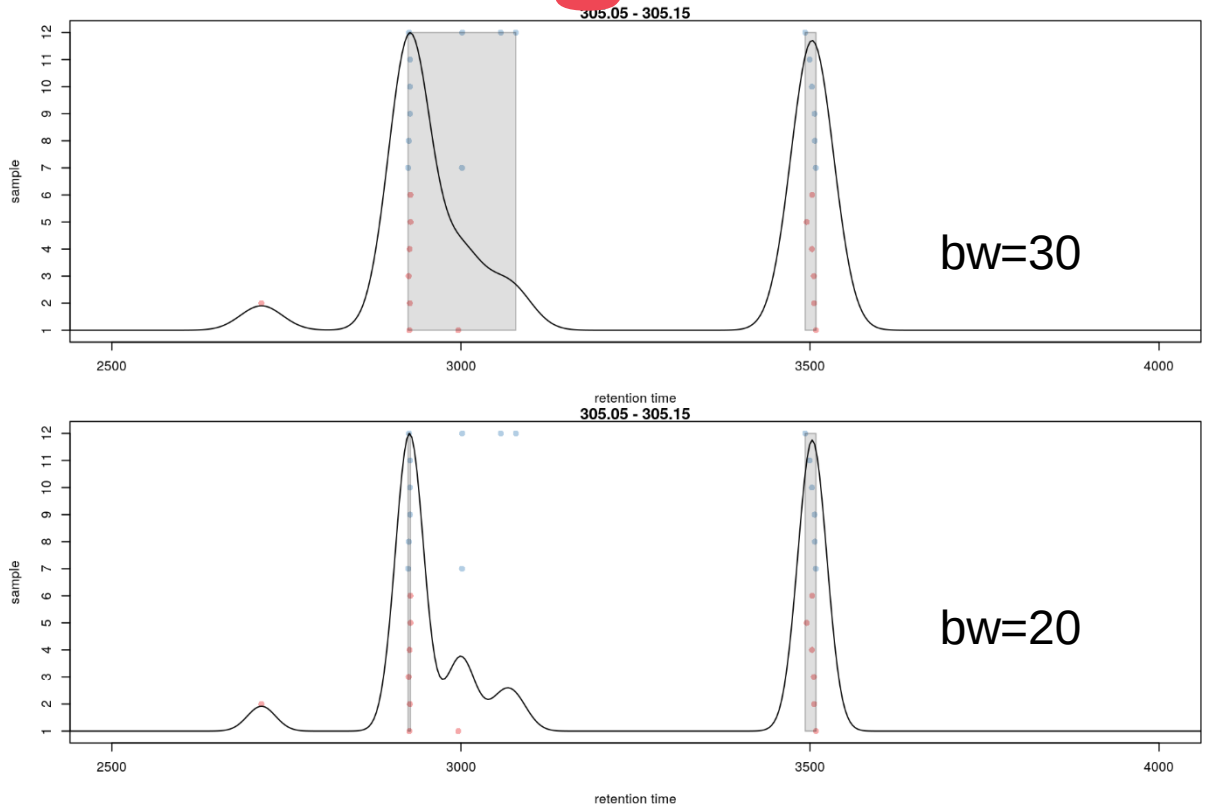
Mass spectrometry data processing with XCMS



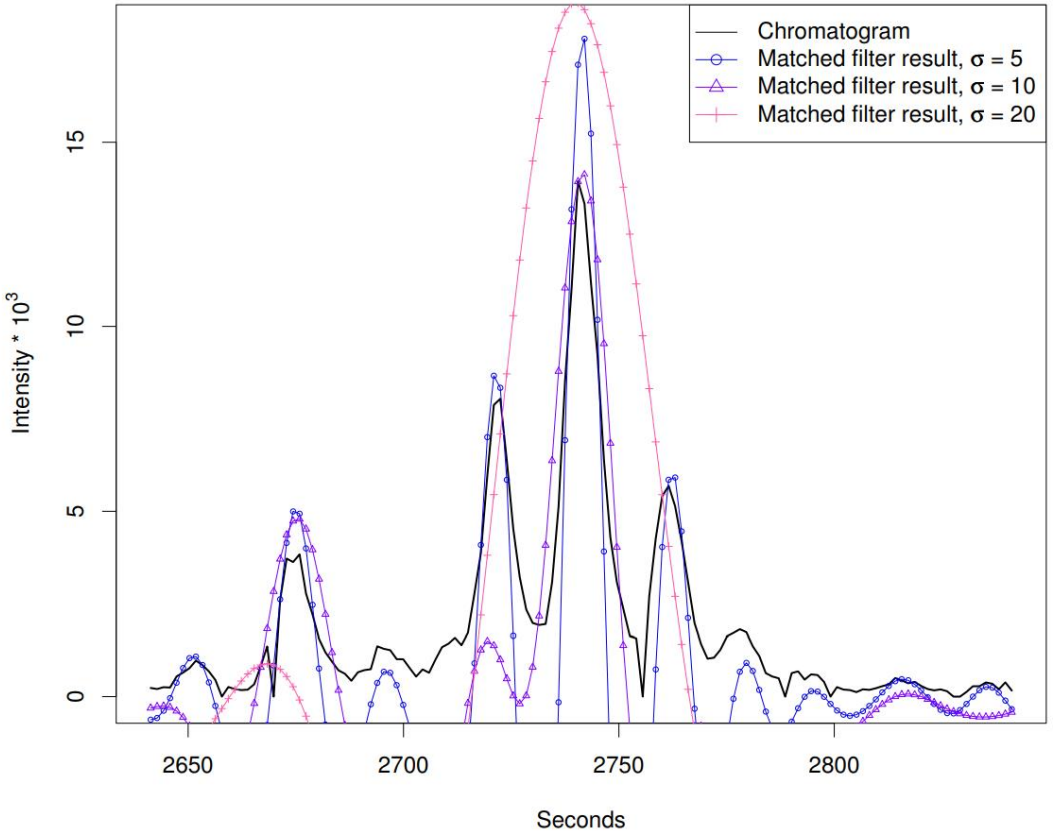
Peak integration with xcms



Peak integration with xcms



Effect of the band width in the distinction of peaks



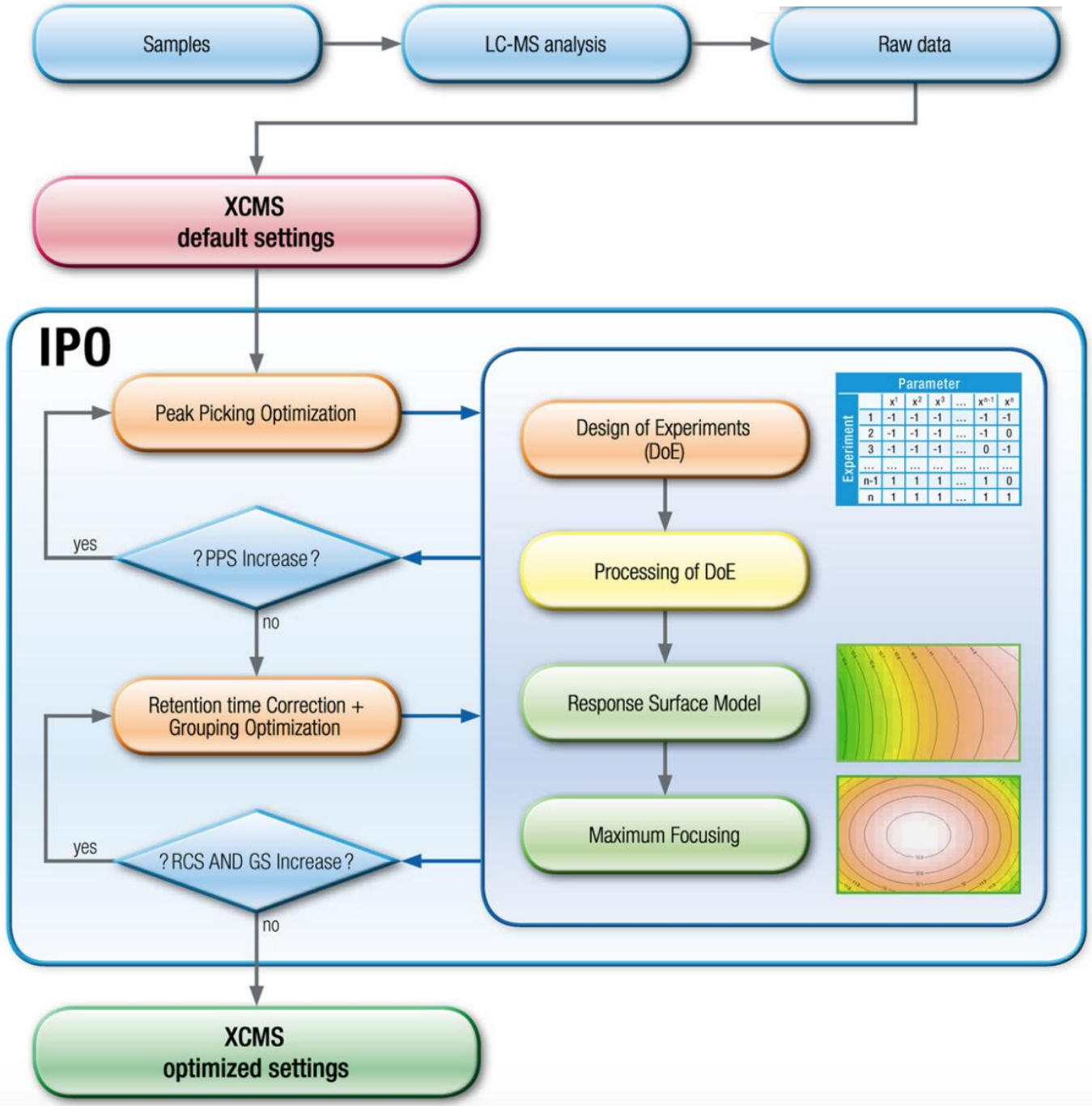
Effect of sigma value width in the distinction of peaks

```
xcmsSet(, method,  ppm , peakwidth, snthresh)
group (, method , bw, mzwid, minfrac, minsamp)
retcor(, method="loess", plottype, span)
```

Automated optimization of XCMS parameters

Which samples to use?

QCs or individual sample across your different conditions ?



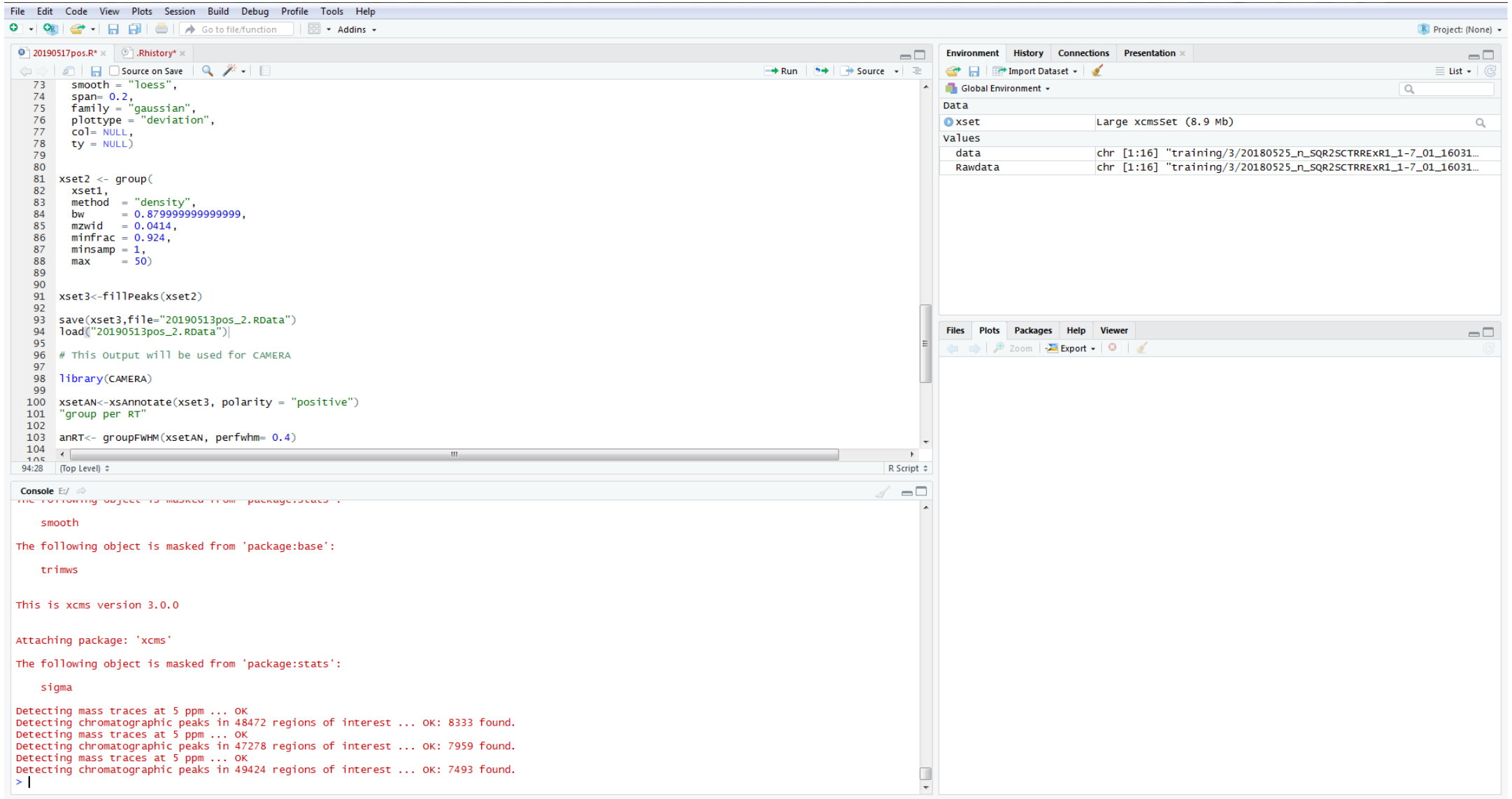
Automated optimization of XCMS parameters

- Peak picking Parameters optimization

- Retention time and grouping optimization

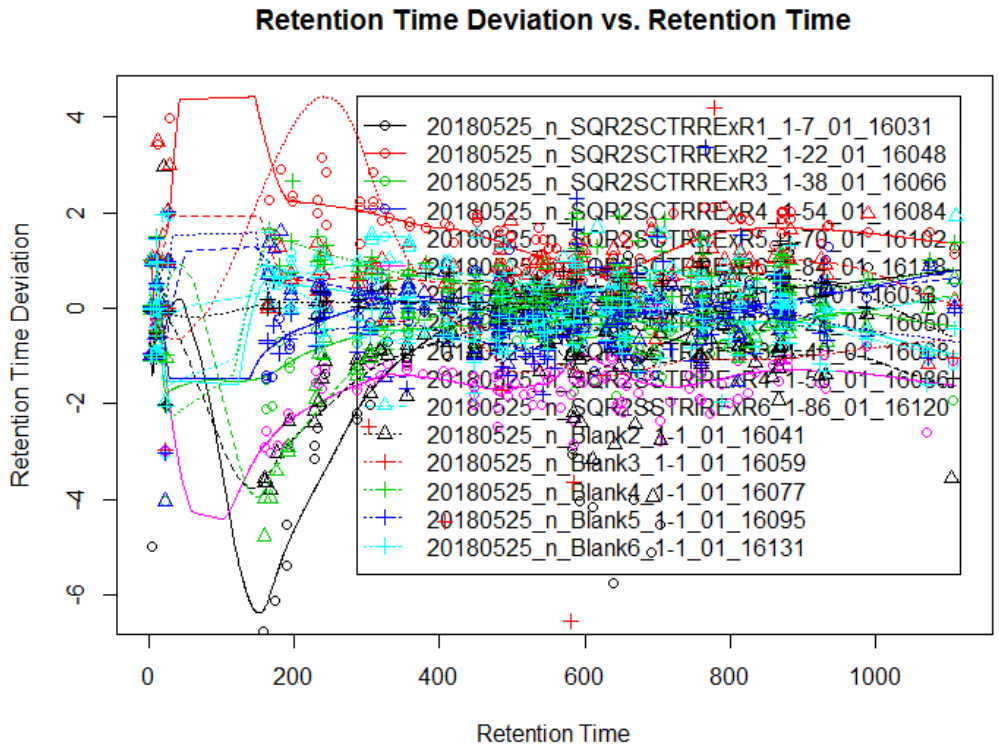
XCMS method	Parameters
xcmsSet(method = 'centWave')	min peakwidth, max peakwidth, ppm, mzdiff
xcmsSet(method = 'matchedFilter')	fwhm, step, steps, snthresh, mzdiff
retcor(method = 'obiwarp')	profStep, gapInit, gapExtend
group(method = 'density')	bw, mzwid, minfrac

Integrating peaks with XCMS

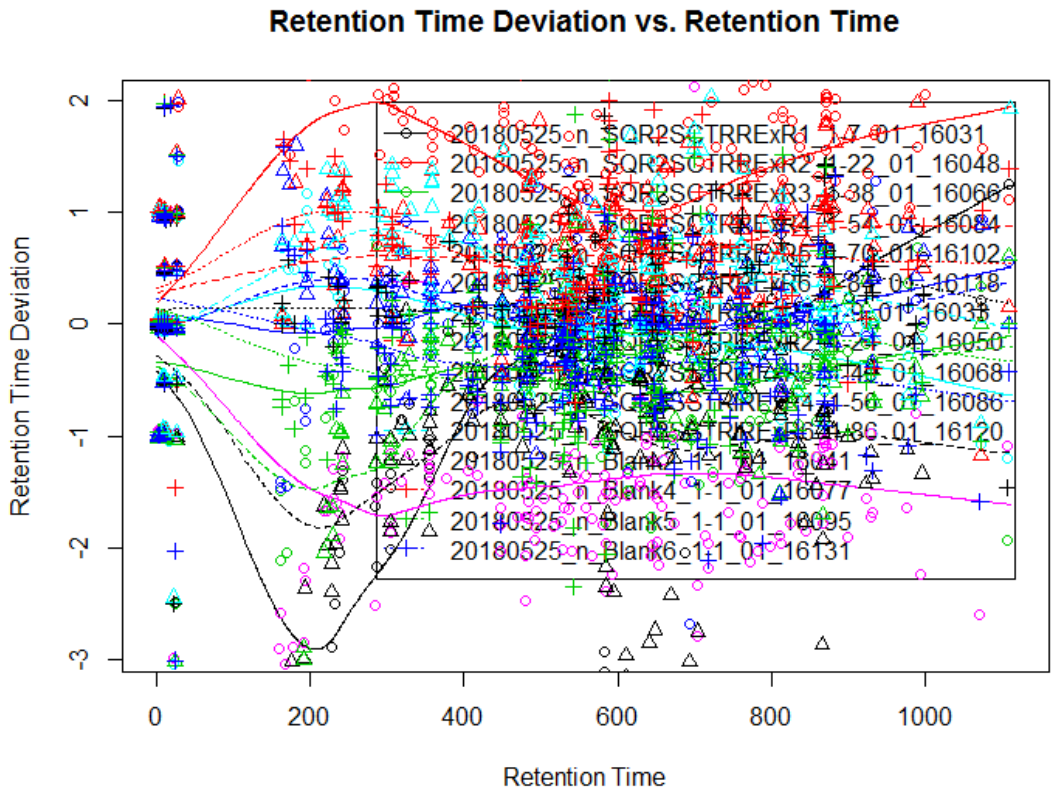


Peaks grouping and retention time correction

With blank 3



Without blank 3

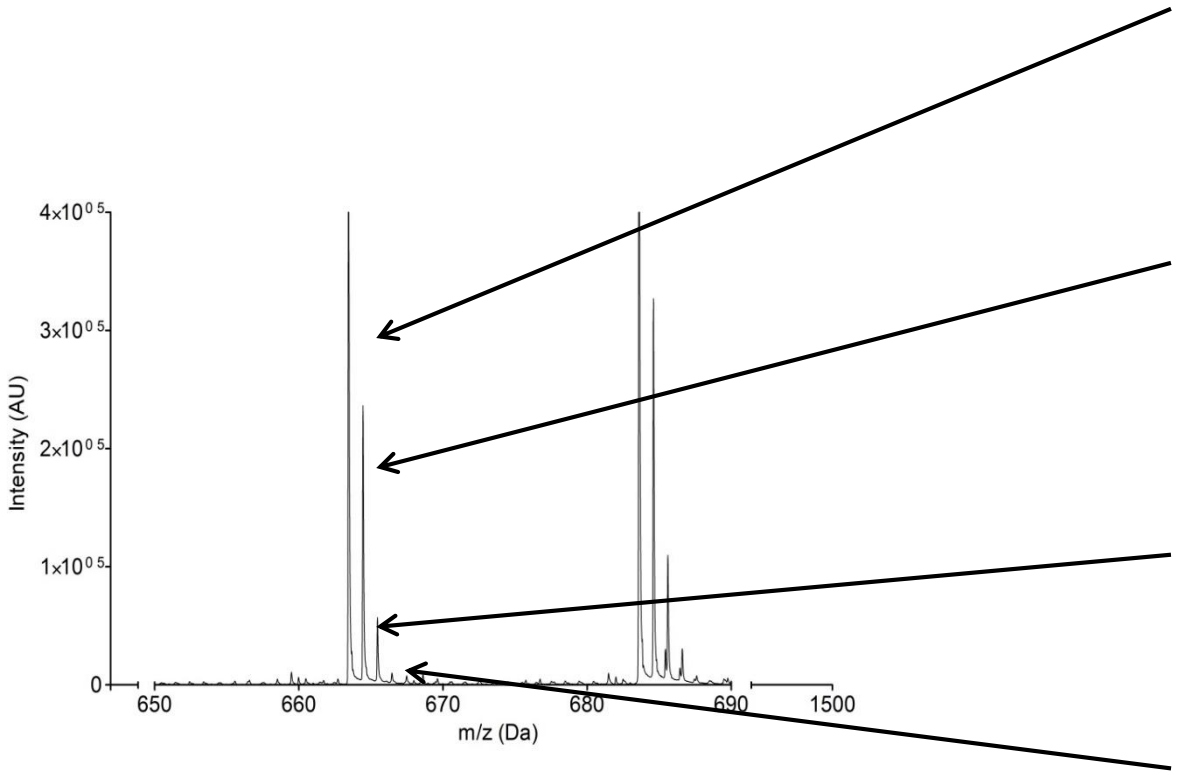


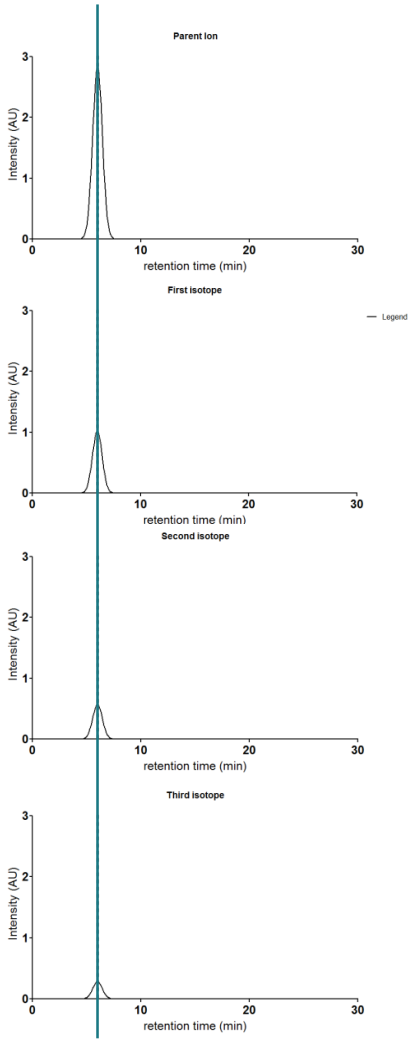
Peak filling

```
> NF[20:30,c(1, 4, 11:17)]
          mz         rt X20180525_n_SQR2SCTRREXR1_1.7_01_16031 X20180525_n_SQR2SCTRREXR2_1.22_01_16048 X20180525_n_SQR2SCTRREXR3_1.38_01_16066
20  96.96011 287.774727                                    NA                                     NA                              34814.328
21  96.96942 157.536304                                    NA                                     NA                                     NA
22  98.95580   4.749831                                    NA                                     NA                                     NA
23  99.92561   2.553982                                    NA                              154676.160                                    NA
24 100.93327  10.538675                              11624.364                                     NA                               5514.480
25 105.03429 194.506980                              11866.985                              10940.163                               5322.276
26 106.00448   2.110093                             193953.852                                     NA                                     NA
27 107.05005 194.546324                              37334.904                              31172.820                              30155.342
28 107.05006  22.219764                                    NA                               7640.352                               9763.857
29 108.05331 194.066886                               3146.069                                     NA                               2208.936
30 110.97551   3.565727                                    NA                              13405.267                                     NA
   X20180525_n_SQR2SCTRREXR4_1.54_01_16084 X20180525_n_SQR2SCTRREXR5_1.70_01_16102 X20180525_n_SQR2SCTRREXR6_1.84_01_16118
20                               44551.654                                      NA                               29900.289
21                                      NA                                      NA                                      NA
22                               28915.056                                      NA                                      NA
23                                      NA                                      NA                                      NA
24                                5559.549                                      NA                                3784.212
```

```
> F[20:30,c(1, 4, 11:17)]
          mz         rt X20180525_n_SQR2SCTRREXR1_1.7_01_16031 X20180525_n_SQR2SCTRREXR2_1.22_01_16048 X20180525_n_SQR2SCTRREXR3_1.38_01_16066
20  96.96011 287.774727                              66756.642                              60690.852                              34814.328
21  96.96942 157.536304                             106684.157                              85882.441                              69024.908
22  98.95580   4.749831                              15690.587                              24044.431                              23407.734
23  99.92561   2.553982                                  0.000                             154676.160                                  0.000
24 100.93327  10.538675                              11624.364                              12745.939                               5514.480
25 105.03429 194.506980                              11866.985                              10940.163                               5322.276
26 106.00448   2.110093                             193953.852                                  0.000                                  0.000
27 107.05005 194.546324                              37334.904                              31172.820                              30155.342
28 107.05006  22.219764                              19378.214                               7640.352                               9763.857
29 108.05331 194.066886                               3146.069                               2036.717                               2208.936
30 110.97551   3.565727                                  0.000                              13405.267                                  0.000
   X20180525_n_SQR2SCTRREXR4_1.54_01_16084 X20180525_n_SQR2SCTRREXR5_1.70_01_16102 X20180525_n_SQR2SCTRREXR6_1.84_01_16118 X20180525_n_SQR2SSTRIREXR1_1.9_01_16033
20                               44551.654                              47909.280                              29900.289                              63286.992
21                               82103.497                              61709.321                              50493.616                              90679.990
22                               28915.056                              15292.173                               7929.404                               8359.782
23                              110014.733                                  0.000                                  0.000                                  0.000
24                                5559.549                               8707.655                               3784.212                               6561.432
```

Isotopes and adducts detection with CAMERA





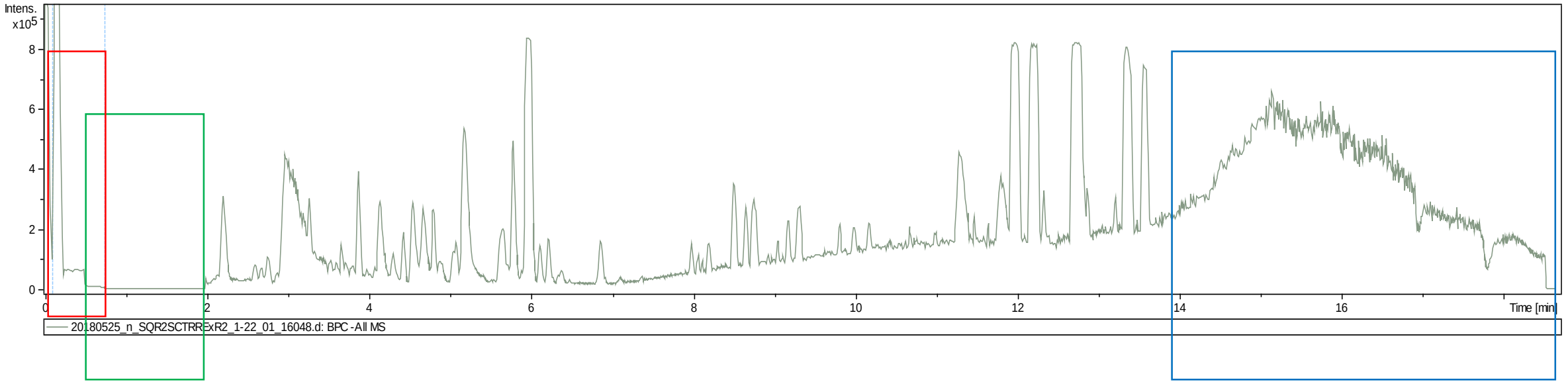
- ppm
- Adducts rules depending on the solvents and ionisation mode

Peak integration output

- Table of intensities associated to each single m/z and retention combination
- Application of corrections to get a final working table

indices	m/z	m/z min	m/z max	RTmin	RT	RTmax	S1	S2	...	Sn	Pcgroup	Adducts	isotopes
1													
2													
...													
n													

Precleaning the output



Trash

Mass calibrant

Column washing with high organic solvent percentage

Corrections

a. Chromatography correction

	Number of rows
Initial table	7344
Chromatography	6198

$120 < RT < 840$

b. Internal/External standard corrections

	Number of rows
Initial table	7344
Chromatography	6198
Blanc correction	4758

I_i sample > 4 I_i blank

Corrections

c. Intra and inter batch (QC) correction

$$CV = \frac{\sigma}{\mu}$$

	Number of rows
Initial table	7344
Chromatography	6198
Blanc correction	4758
QC correction	5 to 10 % peaks

Acknowledgements

Theo Reijmers (Venn Life Sciences)

Adrie Dane (Leiden University)

Jorne Troost (Agilent)

Antoine van Kampen (AMC)

Margriet Hendriks (DSM)