Mass spectrometry and proteomics

Using R/Bioconductor

Laurent Gatto

CSAMA - Bressanone - 25 July 2019

Slides available at: http://bit.ly/20190725csama

On the menu

Morning lecture:

1. Proteomics in R/Bioconductor
2. How does mass spectrometry-based proteomics work?
3. Quantitative proteomics
4. Quantitative proteomics data processing and analysis

Afternoon lab:

- Manipulating MS data (raw and identification data)
- Manipulating quantitative proteomics data
- Data processing and DE

	A	B	C	D	E
1	ENSG-ID	RPKM U2OS	RPKM U251MG	RPKM A431	RPKM RATIO U2C
2	ENSG00000196433	0	0	0	NA
3	ENSG00000166763	0	0	0	NA
4	ENSG00000168781	4.16036	3.877868	2.831423	1.072
5	ENSG00000198746	3.414995	2.182754	2.706358	1.564
6	ENSG00000204131	0.4166259	0	0	NA
7	ENSG00000134184	0	0	0	NA
8	ENSG00000164828	67.3945	53.05943	144.152	1.270
9	ENSG00000126945	9.074128	5.110208	6.99964	1.775
10	ENSG00000185220	1.566899	1.424753	0.9343607	1.099
11	ENSG00000171163	4.878578	14.03779	7.697866	0.347
12	ENSG00000171161	7.504043	15.39958	10.45916	0.487
13	ENSG00000175137	4.635162	14.90939	7.502928	0.310
14	ENSG00000189181	0	0	0	NA
15	ENSG00000177151	0	0	0	NA
16	ENSG00000187701	0	0	0	NA
17	ENSG00000184022	0	0	0	NA
18	ENSG00000183130	0	0	0	NA
19	ENSG00000183310	0	0	0	NA
20	ENSG00000182783	0	0	0	NA
21	ENSG00000188558	0	0	0	NA
22	ENSG00000203661	0	0	0	NA
23	ENSG00000196539	0	0	0	NA
24	ENSG00000196240	0	0	0	NA
25	ENSG00000198104	0	0	0	NA
26	ENSG00000175143	0	0	0	NA
27	ENSG00000196944	0	0	0	NA
28	ENSG00000177174	0	0	0	NA
29	ENSG00000177201	0	0	0	NA
30	ENSG00000177186	0	0	0	NA
31	ENSG00000177212	0	0	0	NA

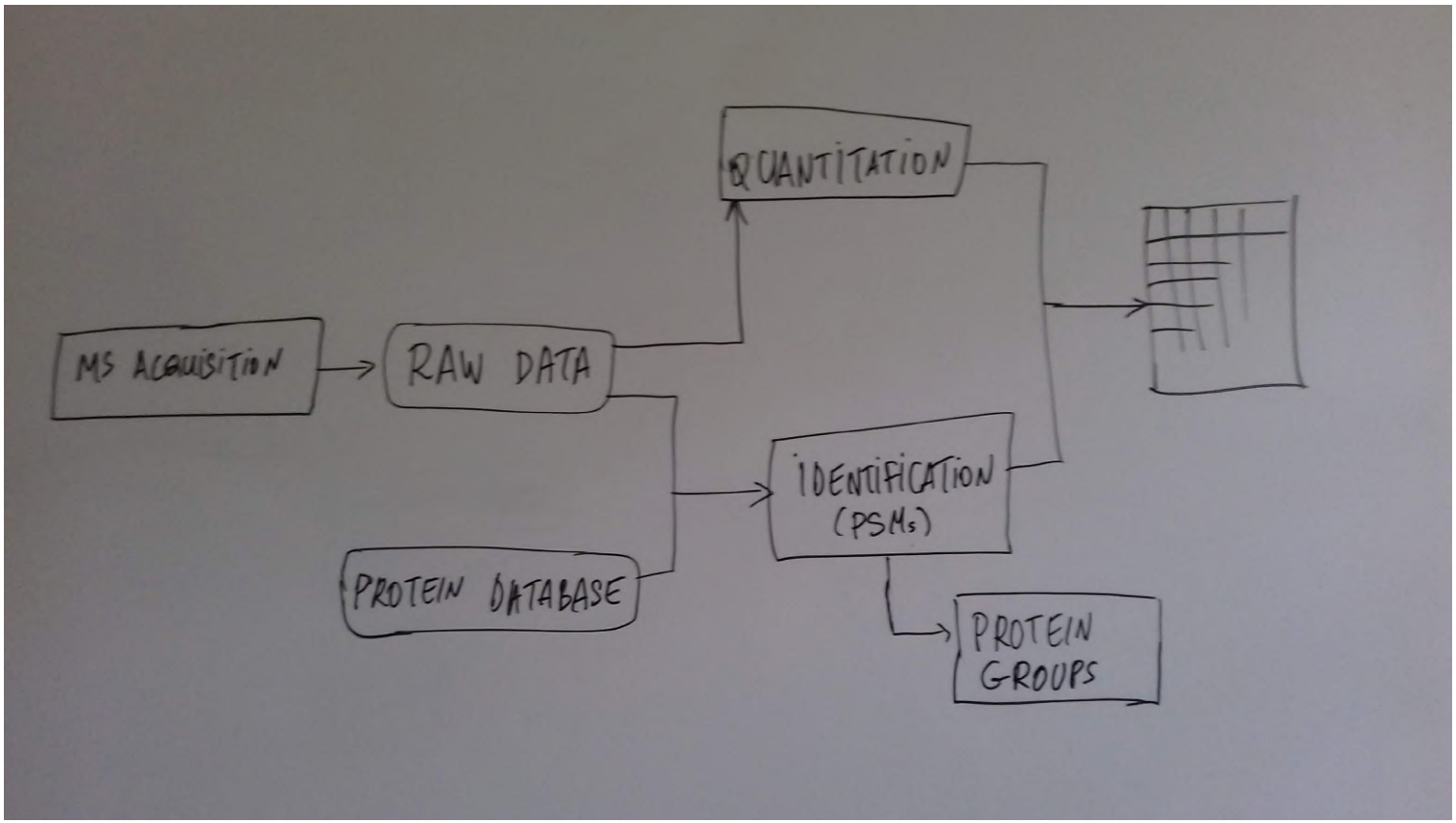
	A	B	C	D	E
1	ENSG-ID	INTENSITY U2O	INTENSITY U251M	INTENSITY A43	SILAC RATIO
2	ENSG00000134184	10499000	340820	704250	0
3	ENSG00000164828	26281000	9406200	28673000	
4	ENSG00000126945	48935000	24463000	40772000	
5	ENSG00000162722	508090	0	0	NA
6	ENSG00000143653	3840200	7028700	5179900	4
7	ENSG00000162851	3545200	3599200	4616700	
8	ENSG00000153187	1091800000	606190000	923910000	0
9	ENSG00000203667	13626000	8258900	13803000	
10	ENSG00000121644	204730	76791	149950	NA
11	ENSG00000035687	66976000	47033000	106990000	
12	ENSG00000117020	1462300	1002000	1272400	
13	ENSG00000143702	4304700	7686900	3976500	
14	ENSG00000203668	4410400	1232100	2063400	3
15	ENSG00000091483	147560000	148660000	137460000	6
16	ENSG00000116984	2563700	1013400	1934000	
17	ENSG00000119285	80282000	39432000	63197000	
18	ENSG00000116977	402720	312940	740150	
19	ENSG00000143669	49550	40389	153270	
20	ENSG00000116957	15326000	11426000	18856000	0
21	ENSG00000152904	1257800	982140	1865100	
22	ENSG00000188739	4634100	3248500	4772200	
23	ENSG00000173726	4458400	3902200	5744800	
24	ENSG00000168264	885950	797480	1048700	
25	ENSG00000168275	1024600	1675900	1146900	
26	ENSG00000135778	11771000	3641700	4445400	1
27	ENSG00000116918	25919000	10251000	13153000	0
28	ENSG00000135766	252780	296600	394150	
29	ENSG00000116903	2575100	1273000	2282500	0
30	ENSG00000119280	1411300	150880	521110	5
31	ENSG00000099977	172740000	49442000	180810000	

1. Proteomics and mass spectrometry packages, questions and workflow in Bioconductor.

2. How does mass spectrometry work?

(applies to proteomics and metabolomics)

Overview


MS ACQUISITION
RAW DATA
QUANTITATION
IDENTIFICATION
(PSMs)
PROTEIN DATABASE
PROTEIN
GROUPS

How does MS work?

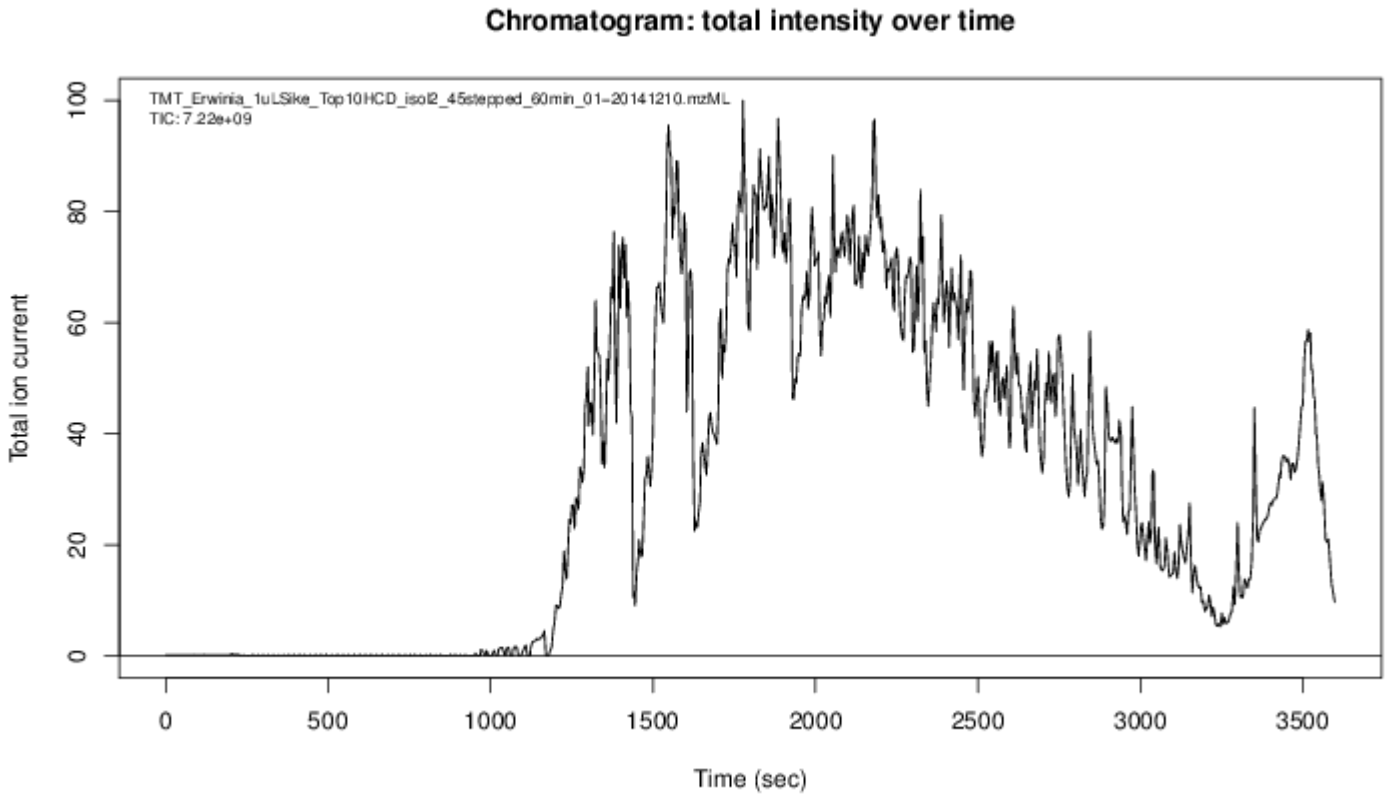
1. Digestion of proteins into peptides - as will become clear later, the features we measure in shotgun (or bottom-up) *proteomics* are peptides, **not** proteins.

2. On-line liquid chromatography (LC-MS)

3. Mass spectrometry (MS) is a technology that **separates** charged molecules (ions, peptides) based on their mass to charge ratio (M/Z).

Chromatography

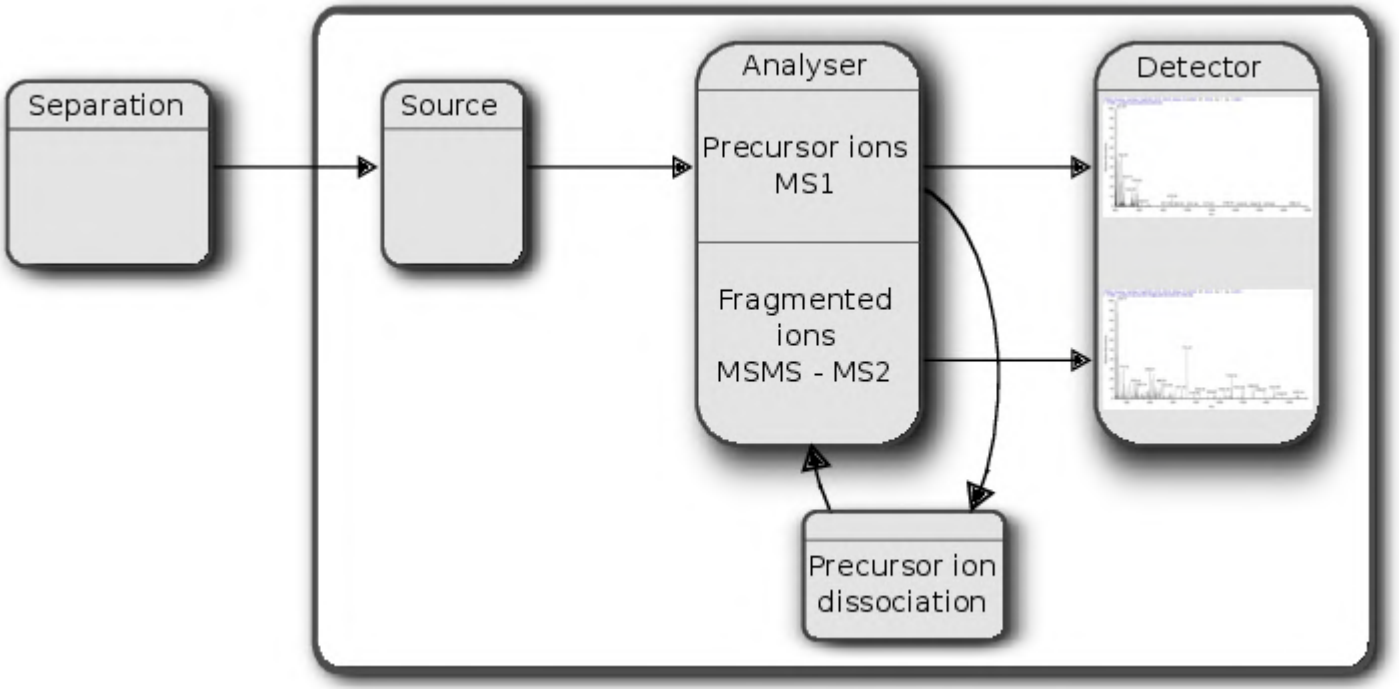
MS is generally coupled to chromatography (liquid LC, but can also be gas-based GC). The time an analytes takes to elute from the chromatography column is the **retention time**.

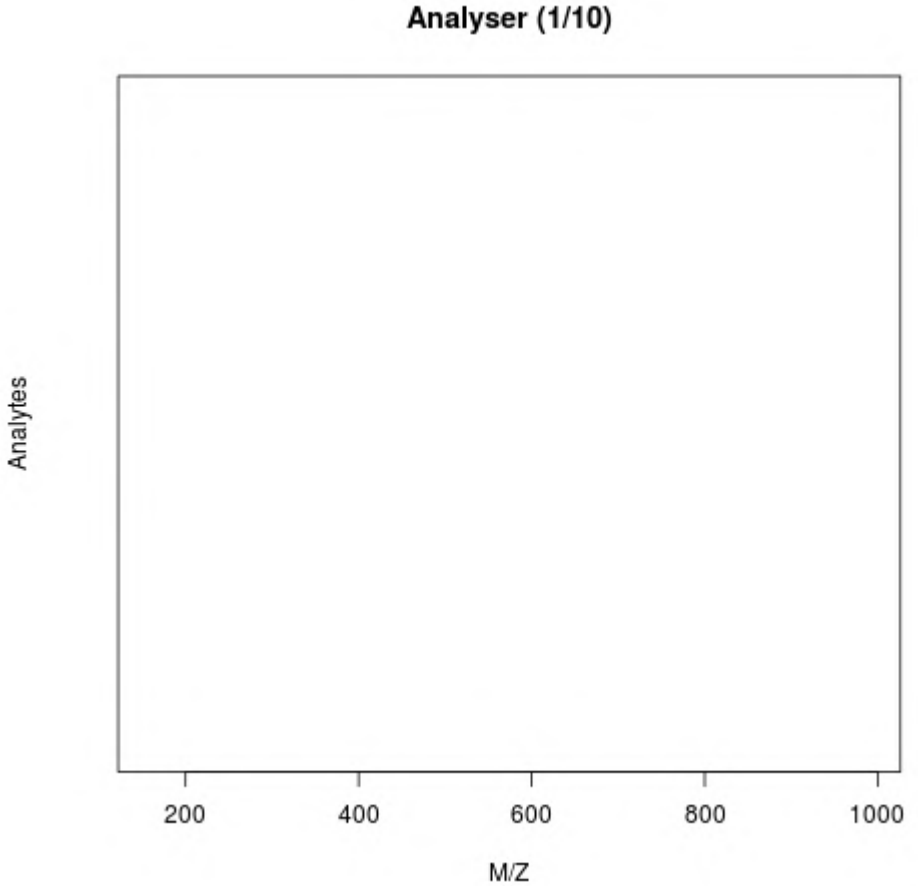


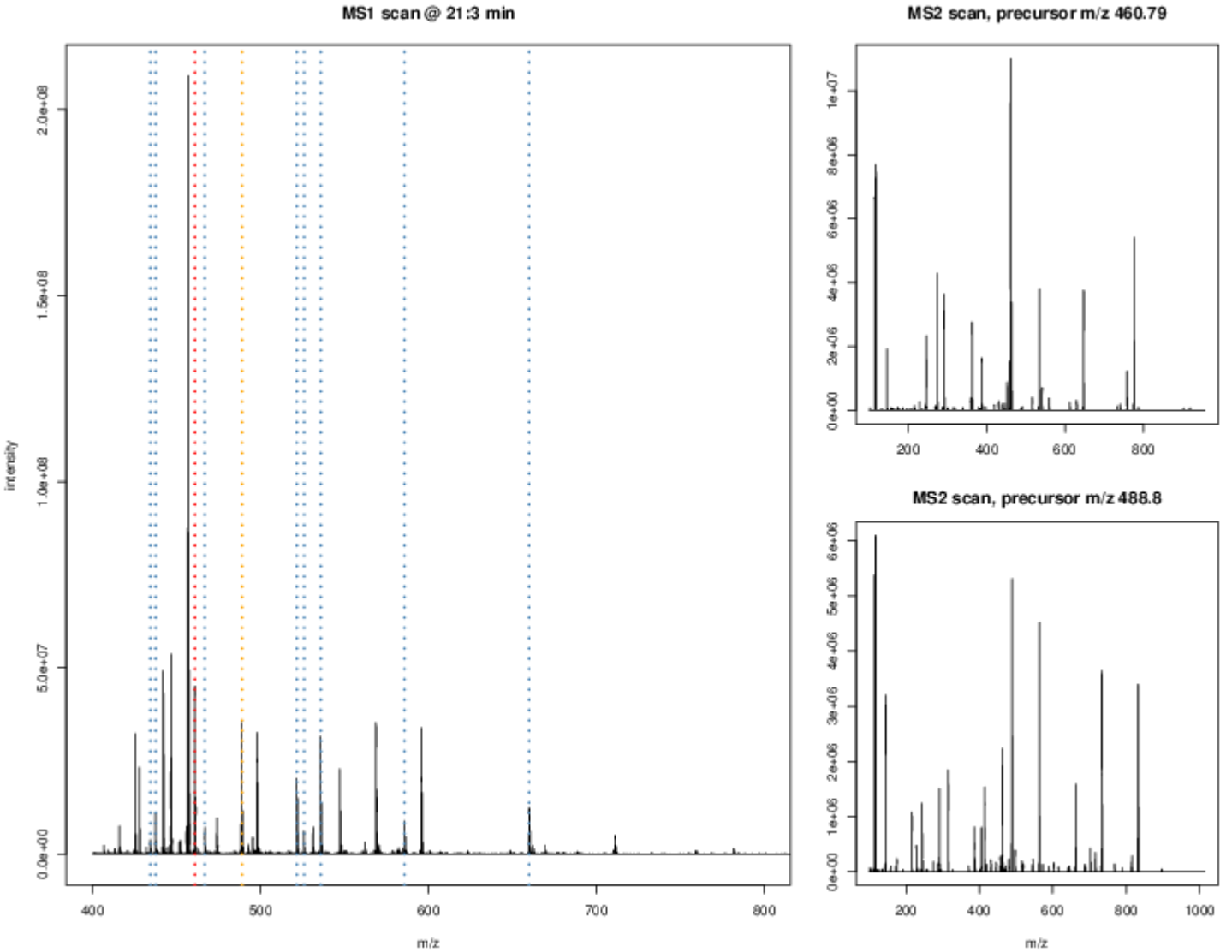
An mass spectrometer is composed of three components:

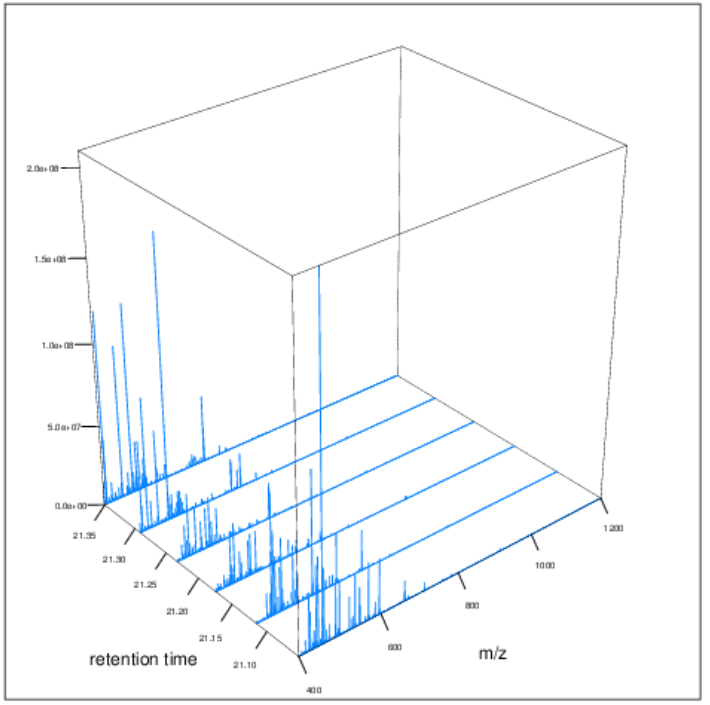
1. The *source*, that ionises the molecules: examples are Matrix-assisted laser desorption/ionisation (MALDI) or electrospray ionisation (ESI).
2. The *analyser*, that separates the ions: Time of flight (TOF) or Orbitrap.
3. The *detector* that quantifies the ions.

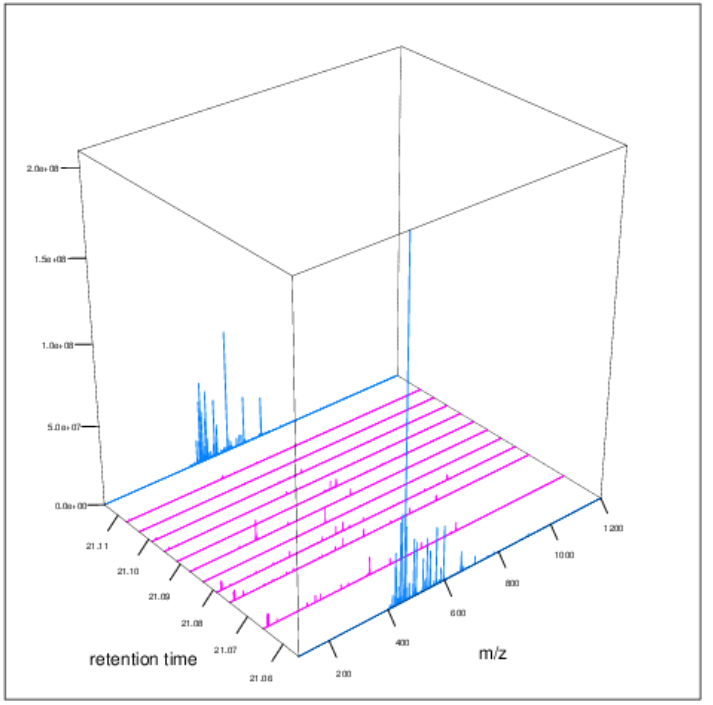
Ions typically go through that cylce at least twice (MS2, tandem MS, or MSMS). Before the second cycle, individual *precursor* ions a selected and broken into *fragment* ions.


Separation
Source
Analyser
Precursor ions
MS1
Fragmented
ions
MSMS - MS2
Detector
Precursor ion
dissociation

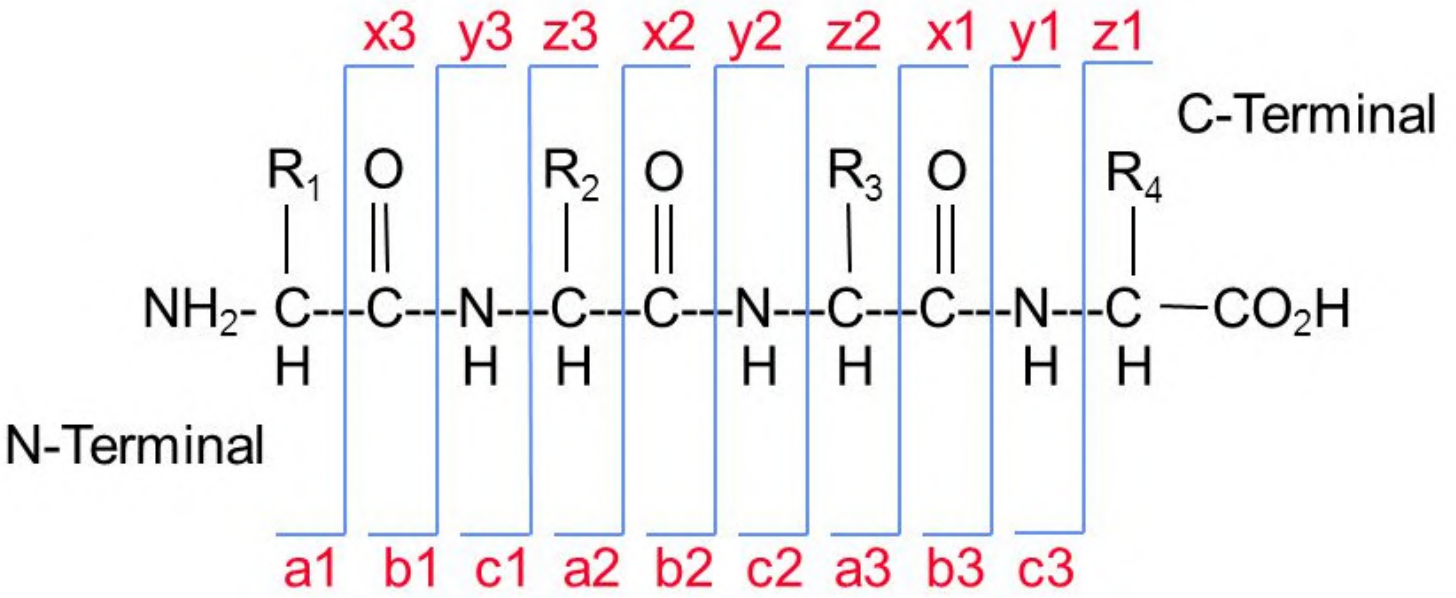

Analyser (1/10)
Analytes
200
400
600
800
1000
M/Z


MS1 scan @ 21:3 min
intensity
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
400
500
600
700
800
m/z
MS2 scan, precursor m/z 460.79
0e+00
2e+06
4e+06
6e+06
8e+06
1e+07
200
400
600
800
MS2 scan, precursor m/z 488.8
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
6e+06
200
400
600
800
1000
m/z


2.0e+08
1.5e+08
1.0e+08
5.0e+07
0.0e+00
21.35
21.30
21.25
21.20
21.15
21.10
retention time
400
600
800
1000
1200
m/z


2.0e+08
1.5e+08
1.0e+08
5.0e+07
0.0e+00
21.11
21.10
21.09
21.08
21.07
21.06
retention time
200
400
600
800
1000
1200
m/z

Identification: fragment ions



Biemann, K *Methods Enzymol* (1990) **193** 886-887

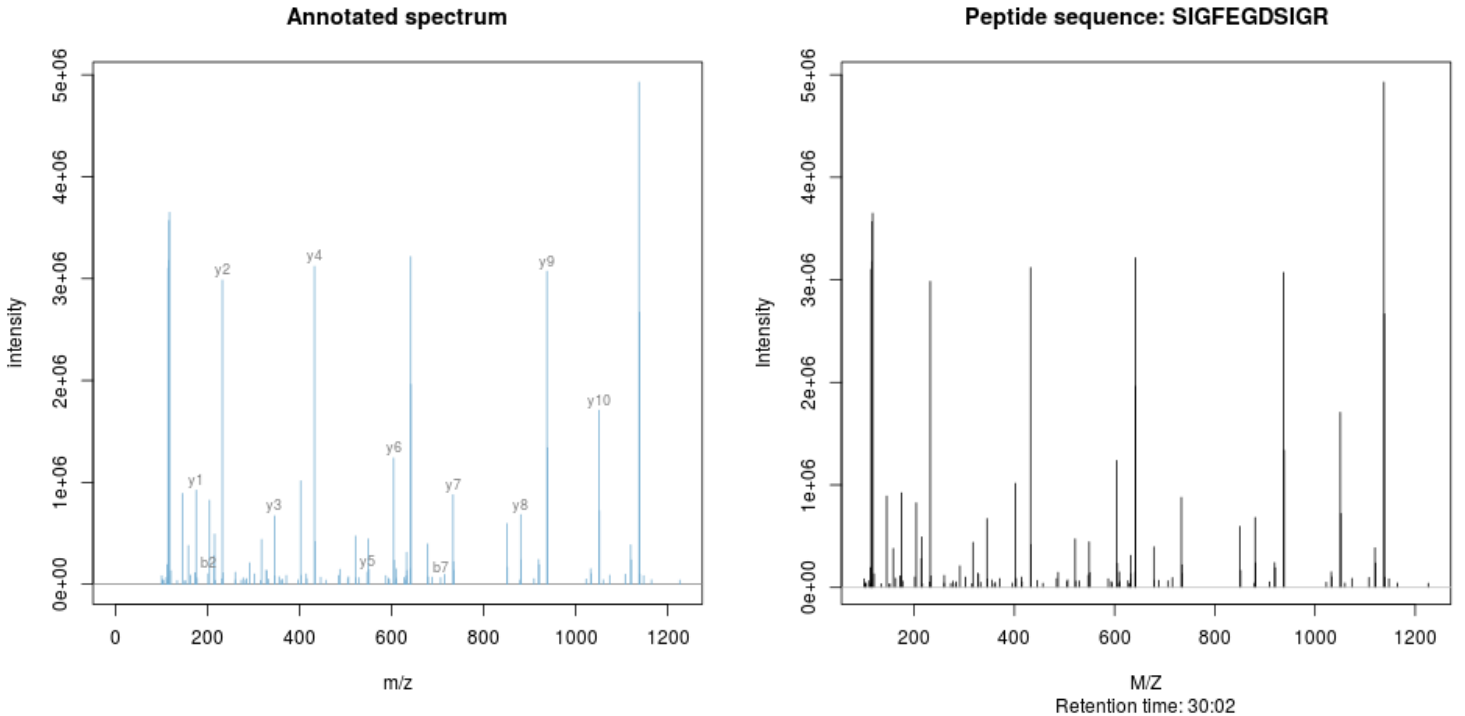
Identification: Peptide-spectrum matching (PSM)

Matching **expected** and *observed* spectra:

```
> MSnbase::calculateFragments("SIGFEGDSIGR")
           mz  ion type pos z        seq
1    88.03931   b1    b   1 1          S
2   201.12337   b2    b   2 1         SI
3   258.14483   b3    b   3 1        SIG
4   405.21324   b4    b   4 1       SIGF
5   534.25583   b5    b   5 1      SIGFE
6   591.27729   b6    b   6 1     SIGFEG
7   706.30423   b7    b   7 1    SIGFEGD
8   793.33626   b8    b   8 1   SIGFEGDS
9   906.42032   b9    b   9 1  SIGFEGDSI
10  963.44178  b10    b  10 1 SIGFEGDSIG
11  175.11895   y1    y   1 1          R
12  232.14041   y2    y   2 1         GR
13  345.22447   y3    y   3 1        IGR
14  432.25650   y4    y   4 1       SIGR
15  547.28344   y5    y   5 1      DSIGR
16  604.30490   y6    y   6 1     GDSIGR
 [ reached getOption("max.print") -- omitted 16 rows ]
```

Identification: Peptide-spectrum matching (PSM)

Matching *expected* and **observed** spectra:



Identification: database

UniProt

Proteomes | Advanced | Search

BLAST | Align | Retrieve/ID mapping | Peptide search | Help | Contact

Proteomes - Homo sapiens (Human)

None

- Overview
- Components
- Publications

Map to

UniProtKB (71,607)

Reviewed (20,336)
Swiss-Prot

Unreviewed (51,271)
TrEMBL

Overview

Status	Reference proteome
Proteins	71,607
Proteome ID	UP000005640
Taxonomy	9606 - Homo sapiens
Last modified	April 5, 2018
Genome assembly and annotation	GCA_000001405.25 from Ensembl

© news.nationalgeographic.com

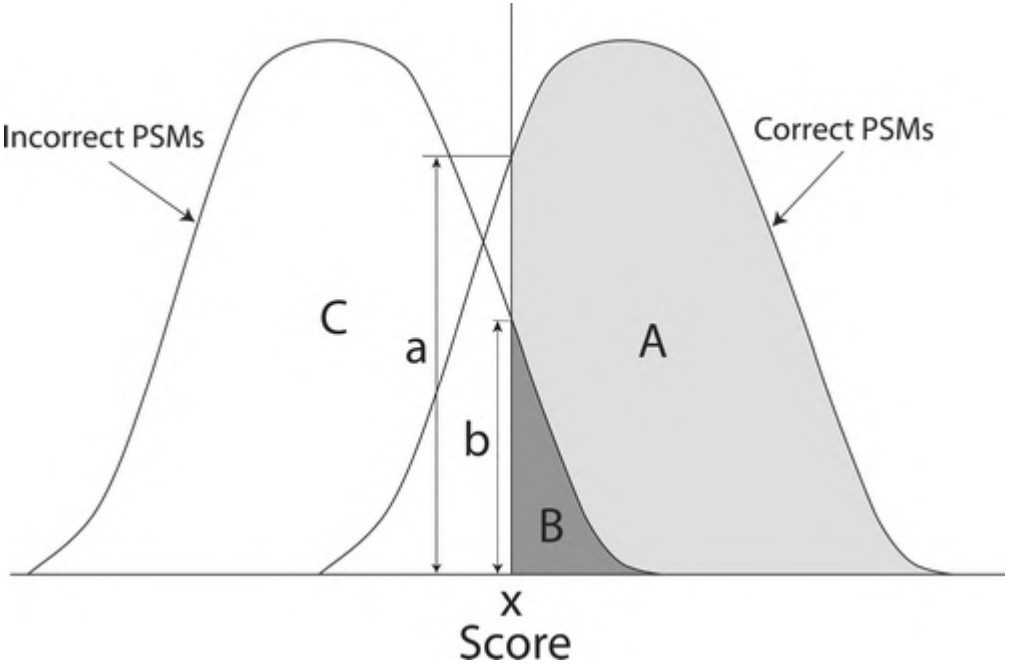
Homo sapiens (Homo sapiens sapiens) or modern humans are the only living species of the evolutionary branch of great apes known as hominids. Divergence of early humans from chimpanzees and gorillas is estimated to have occurred between 4 and 8 million years ago. The genus Homo (Homo habilis) appeared in Africa around 2.3 million years ago and shows the first signs of stone tool usage. The exact lineage of Homo species ie: H. habilis/H. ergaster to H. erectus to H. rhodesiensis/H.heidelbergensis to H. sapiens is still hotly disputed. However, continuing evolution and in particular larger brain size and complexity culminates in Homo sapiens. The first anatomically modern humans appear in the fossil record around 200,000 years ago. Modern humans migrated across the globe essentially as hunter-gatherers until around 12,000 years ago when the practice of agriculture and animal domestication enabled large populations to grow leading to the development of civilizations. Overall life expectancy in Europe is 81 years.

Components

Download | View all proteins

Component name	Genome Accession(s)	Proteins
Chromosome 1	CM000663	5563
Chromosome 2	CM000664	4596
Chromosome 3	CM000665	4122

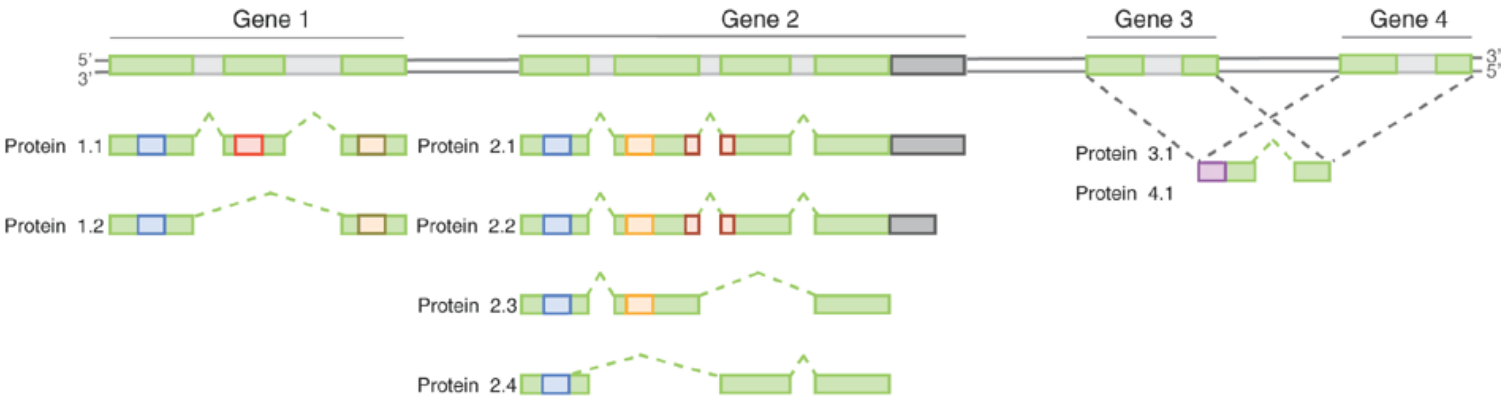
Identification



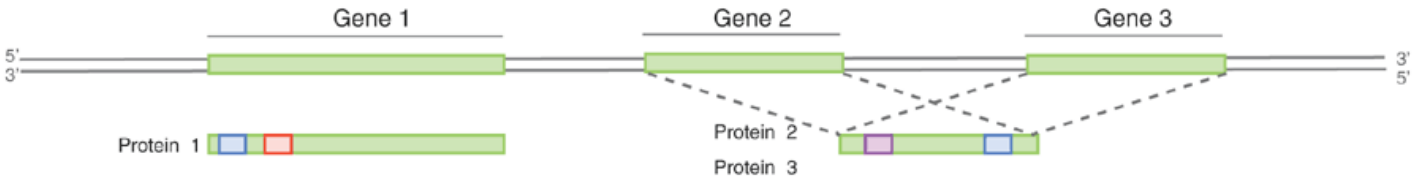
From Käll *et al.* Posterior Error Probabilities and False Discovery Rates: Two Sides of the Same Coin.

Identification: Protein inference

- Keep only reliable peptides
- From these peptides, infer proteins
- If proteins can't be resolved due to shared peptides, merge them into **protein groups** of indistinguishable or non-differentiable proteins.



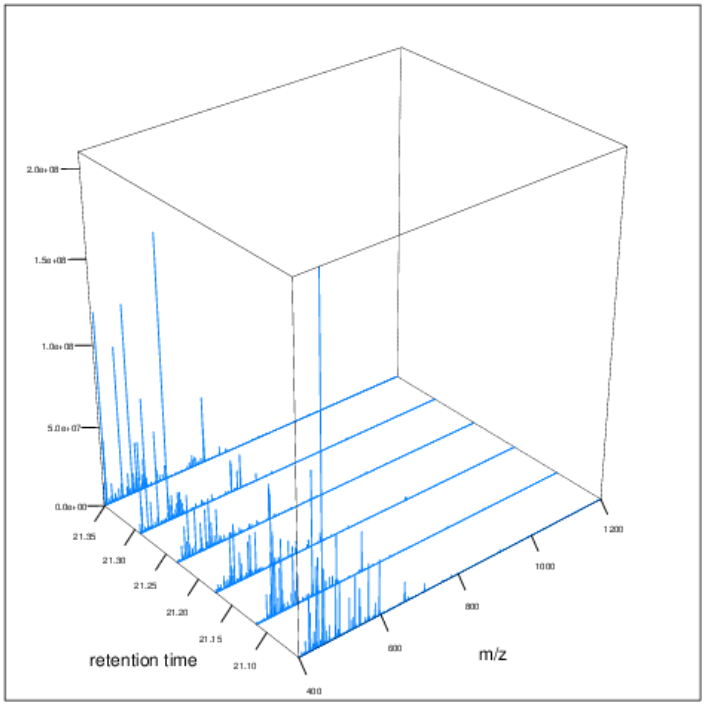
Class	Protein sequence(s)	Protein isoform(s)	Gene(s)
1a	Unambiguous	Unambiguous	Unambiguous
1b	Unambiguous	Ambiguous	Unambiguous
2a	Ambiguous	Ambiguous	Unambiguous
2b	Ambiguous	Ambiguous	Unambiguous
3a	Unambiguous	Ambiguous	Ambiguous
3b	Ambiguous	Ambiguous	Ambiguous

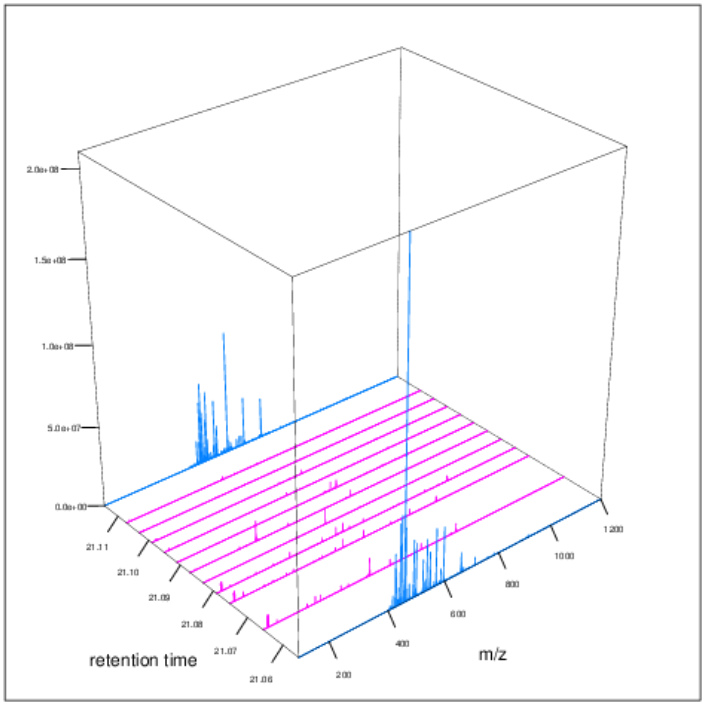


From Qeli and Ahrens (2010).

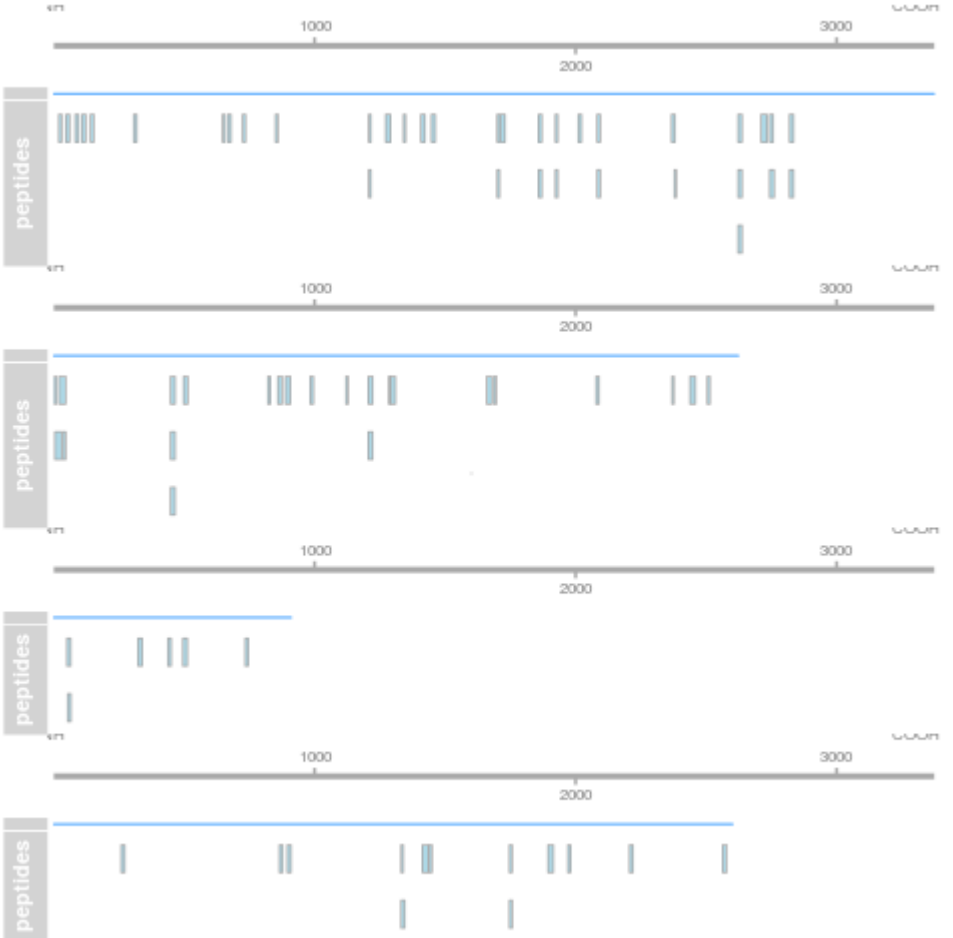
3. Quantitative proteomics

	Label-free	Labelled
MS1	XIC	SILAC, 15N
MS2	Counting	iTRAQ, TMT

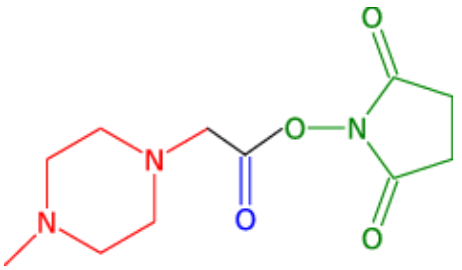


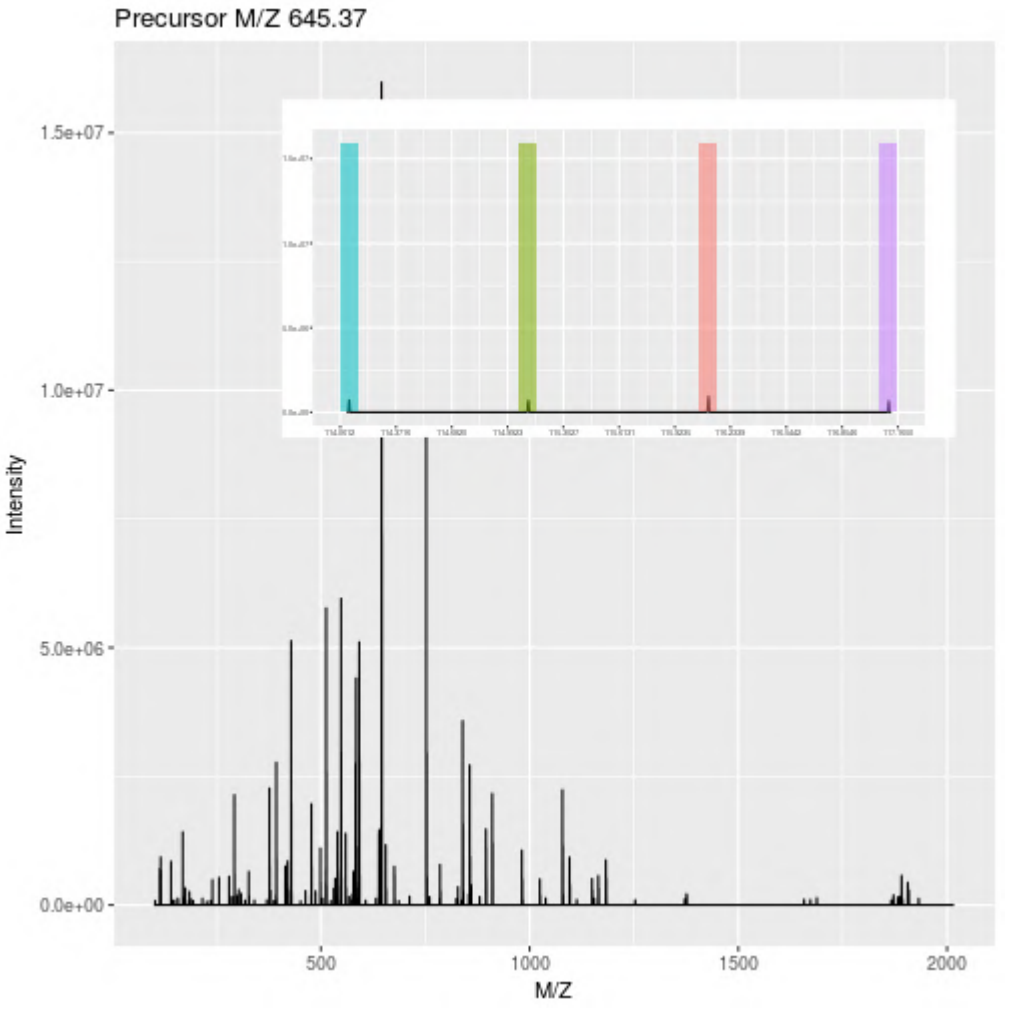


Label-free MS2: Spectral counting

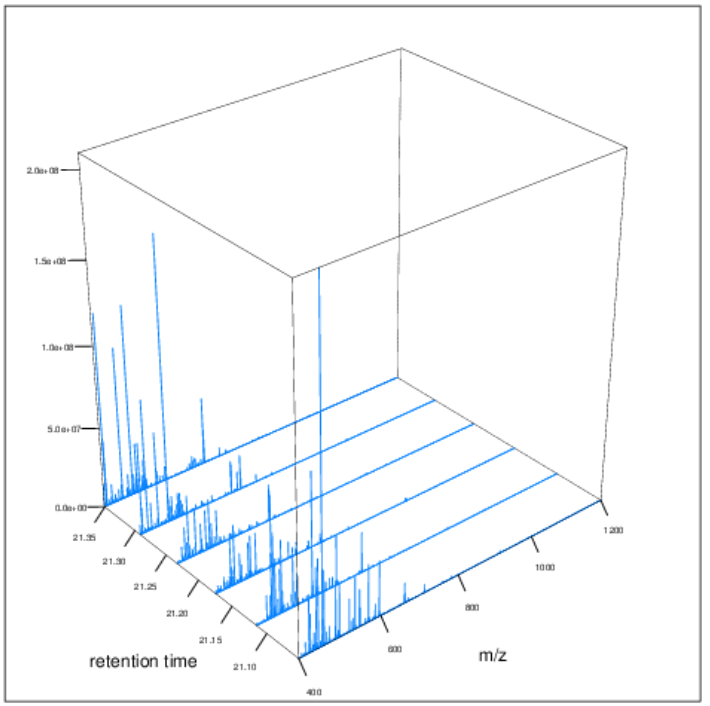


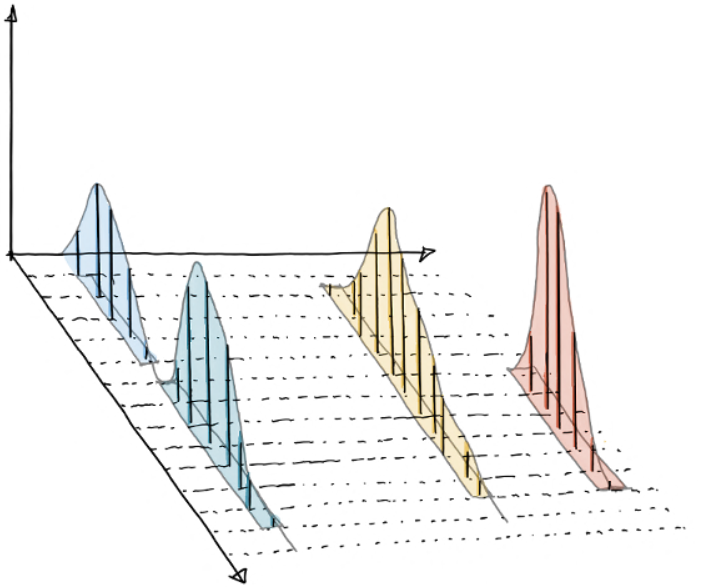
Labelled MS2: Isobaric tagging





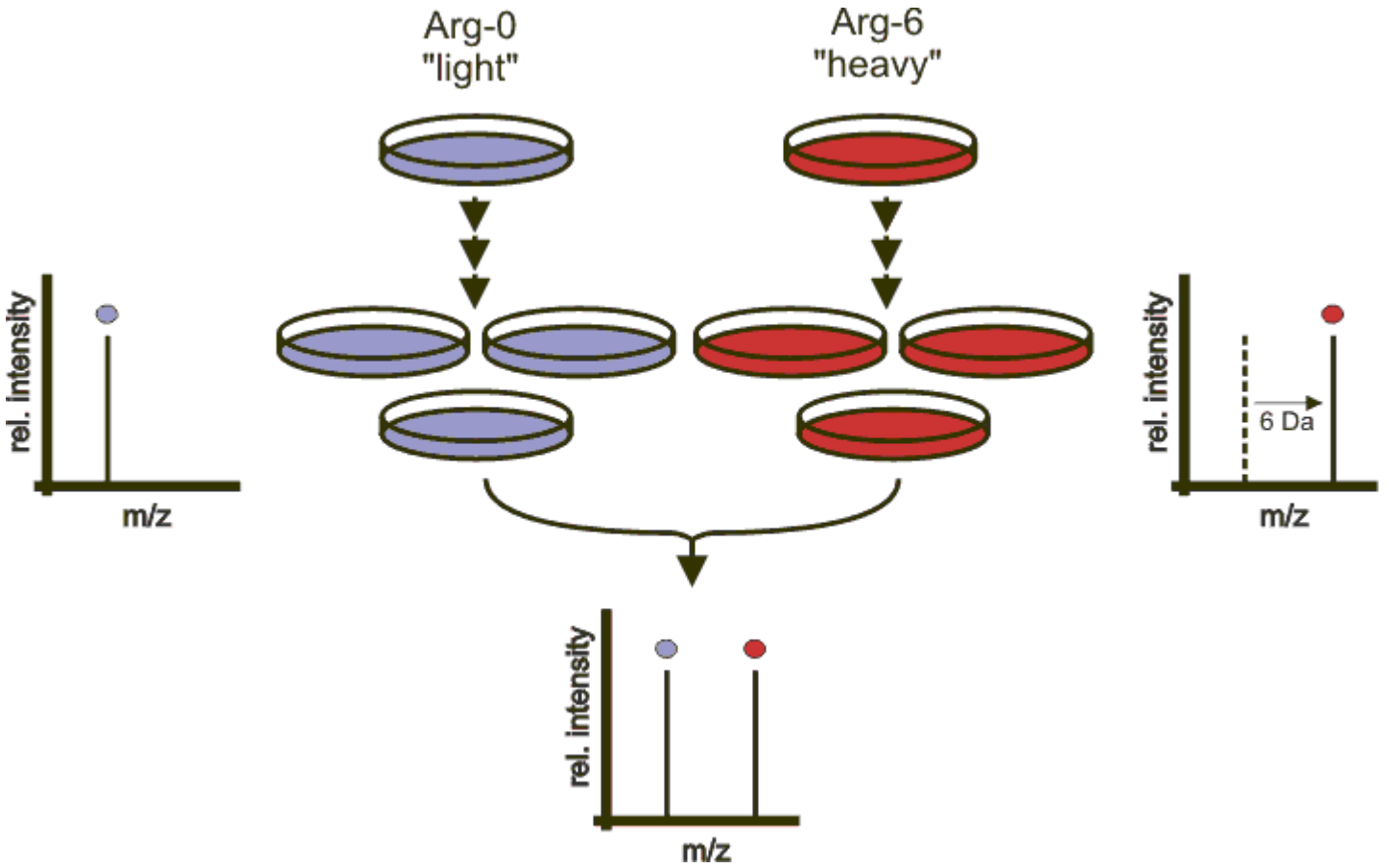
Label-free MS1: extracted ion chromatograms





Credit: Johannes Rainer

Labelled MS1: SILAC



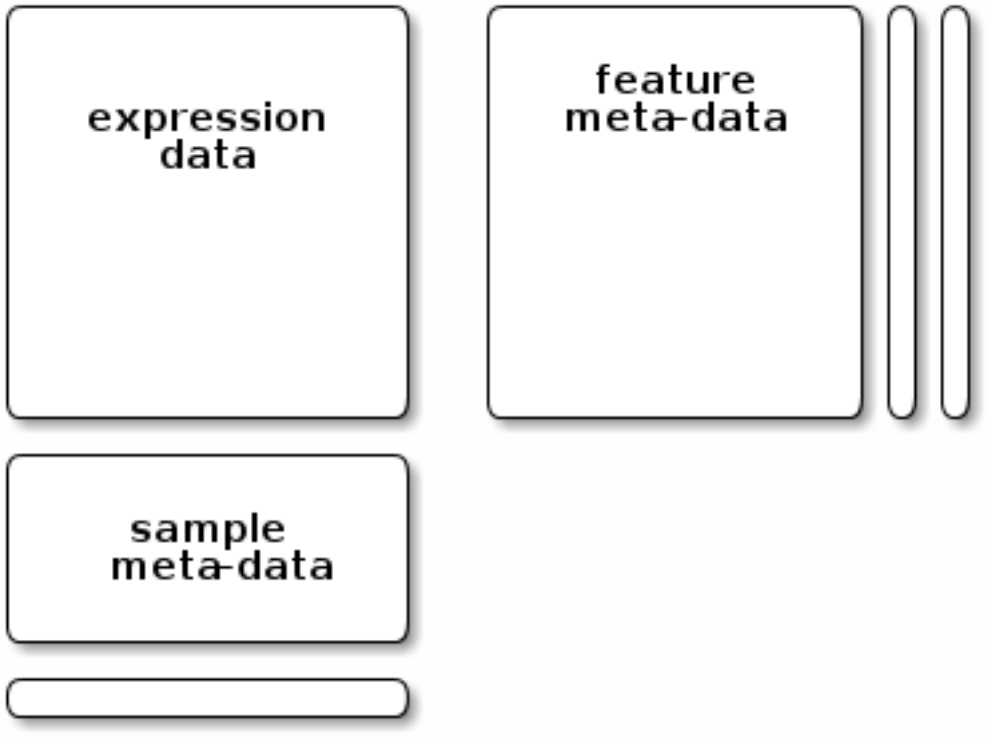
Credit: Wikimedia Commons.

4. Quantitative proteomics data processing and analysis

You will be use the*MSnbase* and *MSqRob* packages during the lab.

Quantitative proteomics data processing

- data processing/cleaning
- **missing values**
- log transformation and normalisation
- **summarisation**
- **differential analysis**



The `MSnSet` structure: expression (accessed with `exprs`), feature (`fData`) and sample (`pData`) metadata.

Missing values

Can appear because

- the feature is missing (due to biology, i.e not at random)
- the feature was missed during the acquisition process (i.e at random)
- mixture thereof

What can one do?

- filter out features (or at least those that have *too many* missing values).
- imputation
- when possible, use a statistical method that accounts for missing values (for example proDA, MSqRob, ...)

Missing values

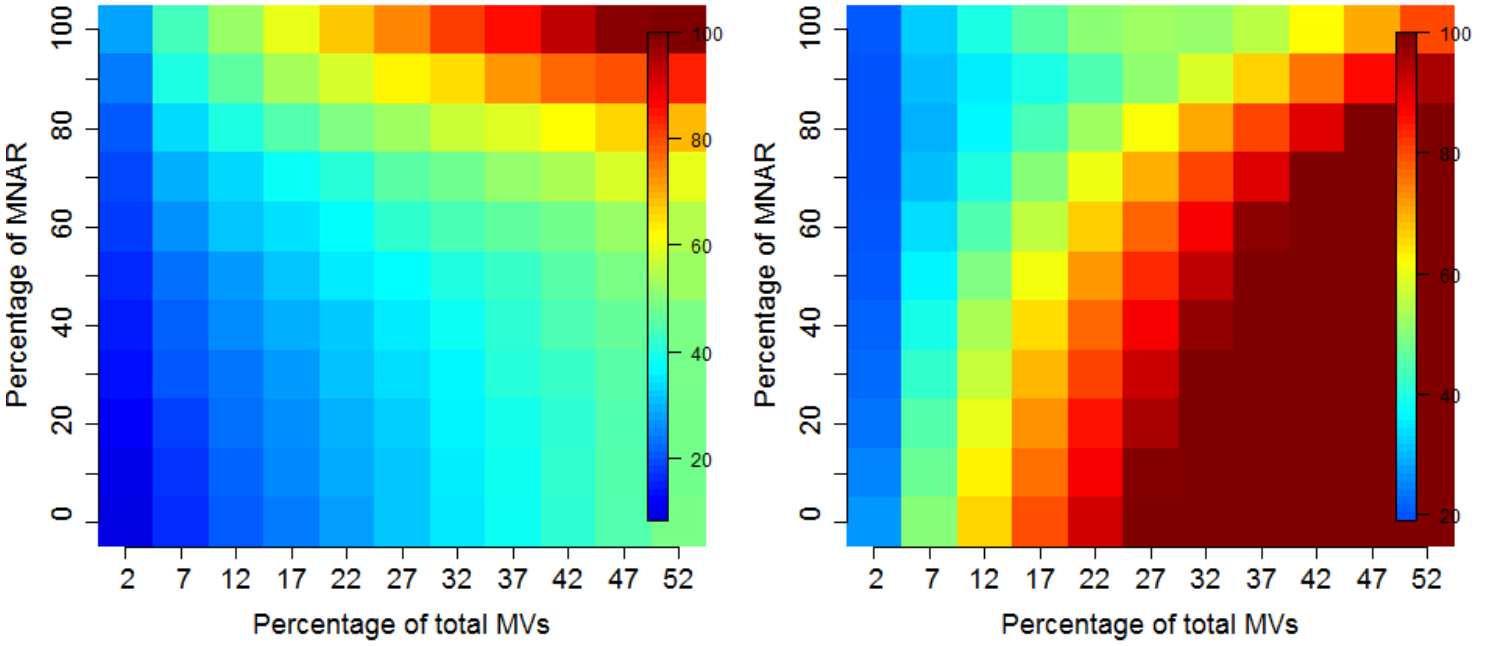
Can appear because

- the feature is missing (due to biology, i.e not at random) - for example `impute(, method = "min")`
- the feature was missed during the acquisition process (i.e at random) - for example `impute(, method = "MLE")`
- mixture thereof - for example `impute(, method = "mixed")` (used in *DEP*)

What can one do?

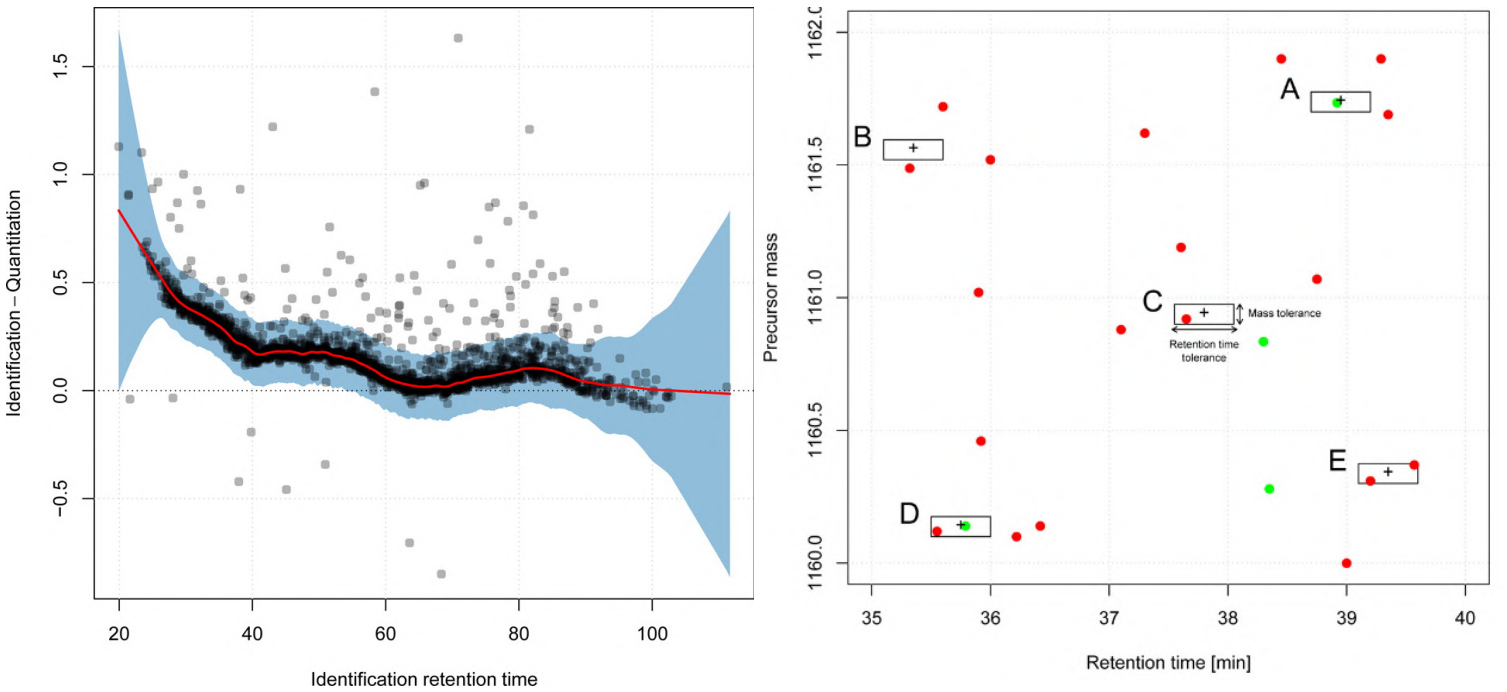
- filter out features (or at least those that have *too many* missing values).
- imputation (`MSnbase::impute` - see above and next slide)
- Feature rescuing (identification transfer, matching between runs)
- when possible, use a statistical method that accounts for missing values (for example proDA, MSqRob, ...)

Imputation



Root-mean-square error (RMSE) observations standard deviation ratio (RSR), KNN and MinDet imputation. Lower (blue) is better. See Lazar *et al.* Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies.

Feature rescuing



From Bond *et al.* Improving Qualitative and Quantitative Performance for MSE-based Label-free Proteomics.

Missing values

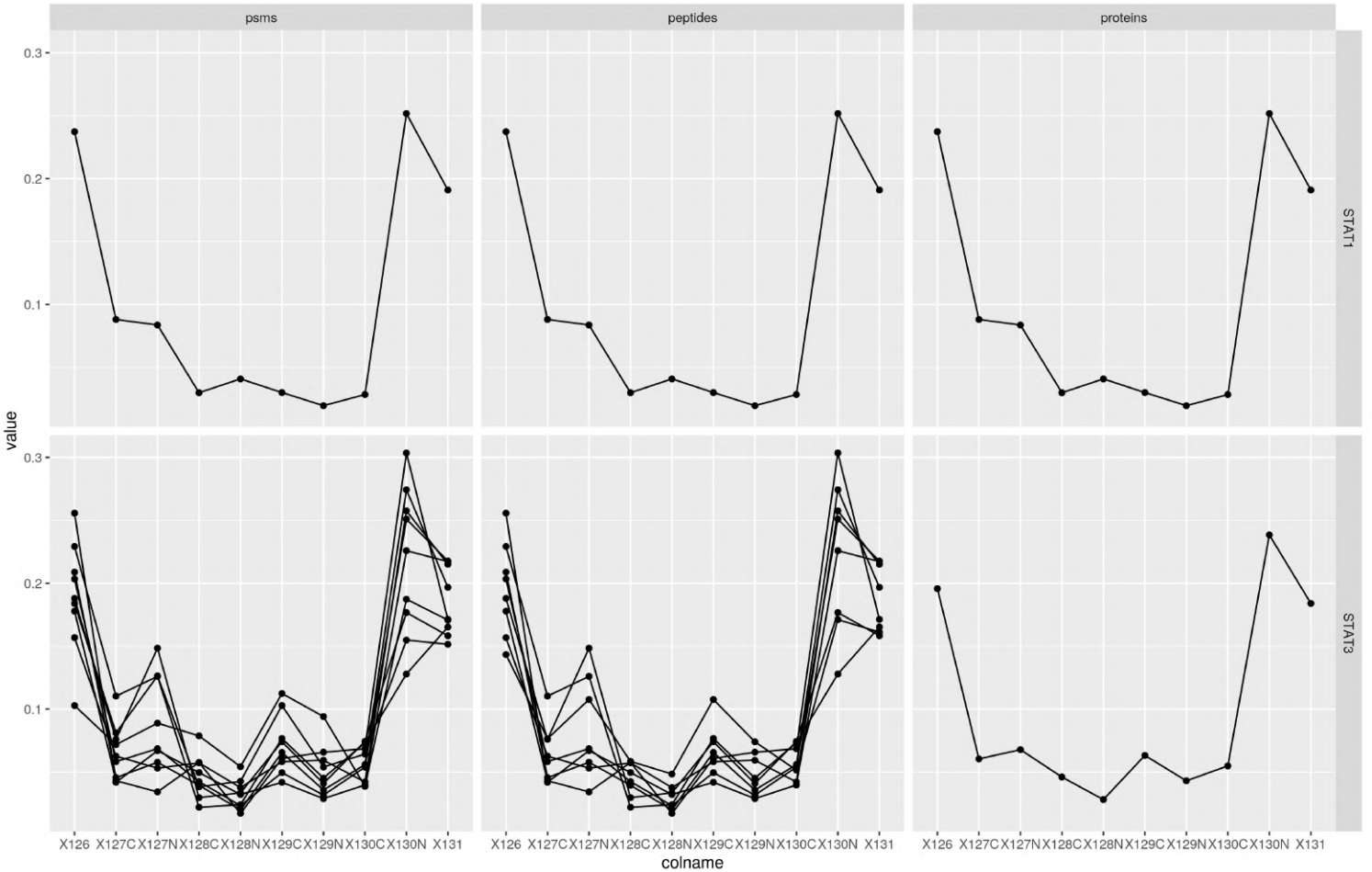
Can appear because

- the feature is missing (due to biology, i.e not at random) - for example `impute(, method = "min")`
- the feature was missed during the acquisition process (i.e at random) - for example `impute(, method = "MLE")`
- mixture thereof - for example `impute(, method = "mixed")` (used in *DEP*)

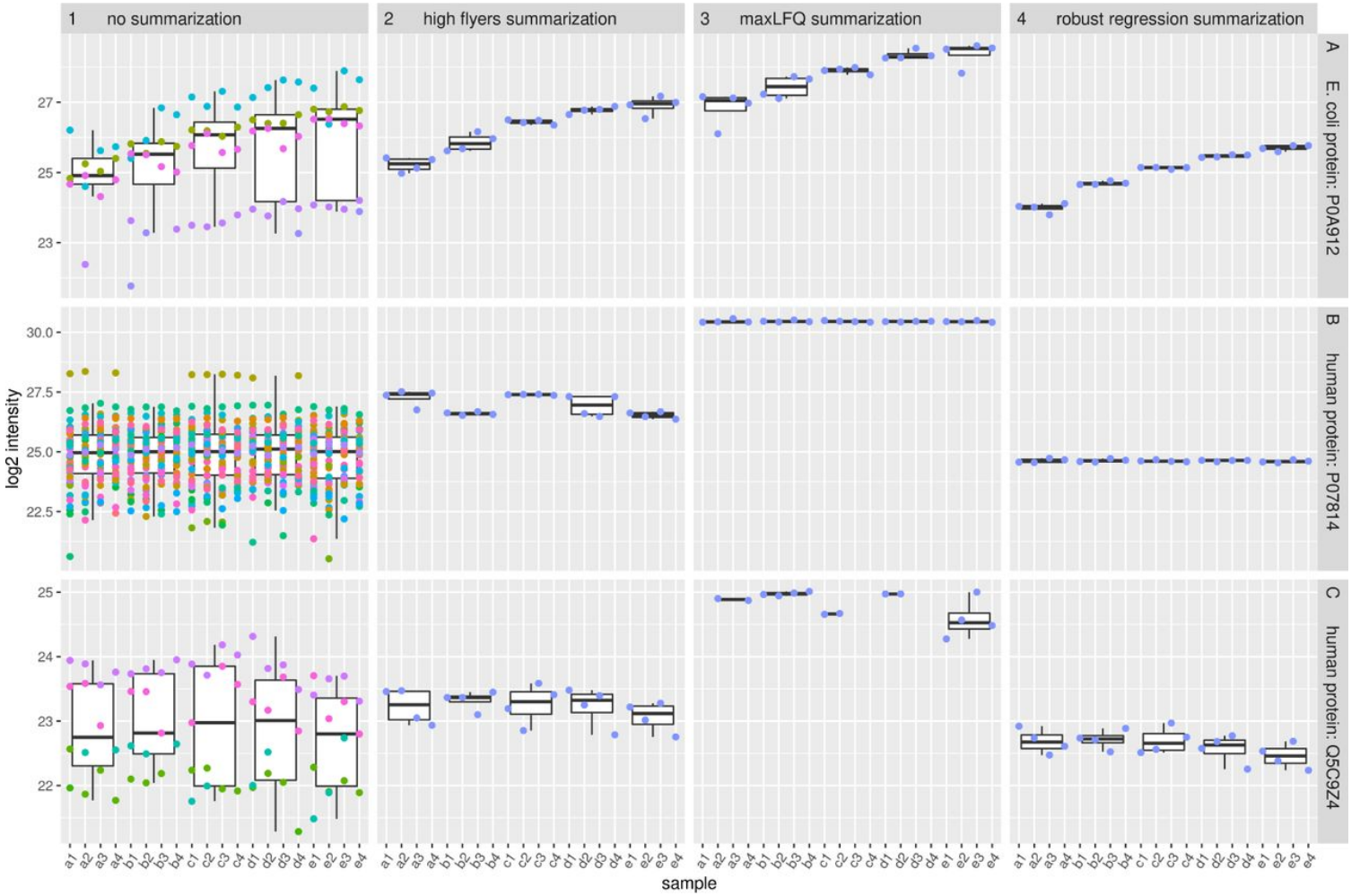
What can one do?

- filter out features (or at least those that have *too many* missing values).
- imputation (`MSnbase::impute` - see above and next slide)
- Feature rescuing (identification transfer, matching between runs)
- when possible, use a statistical method that accounts for missing values (for example proDA, MSqRob, ...)

Summarisation



Examples of aggregations (from the Features package).



Summarisation examples: (1) peptide-level data, (2) mean intensity of the *top three* peptides (Proteus), (3) using pair-wise abundance ratios of shared peptides between samples (MaxQuant) and (4) robust summarisation (MSqRob). From Sticker et al. (2019).

Differential analysis

1. Aggregate normalised peptide intensities of a protein using robust regression with M-estimation using Huber weights:

   $$y_{sp} = \beta_s^{sample} + \beta_p^{pep} + \epsilon_{sp}$$

2. Protein-level inference:

   $$y_{st} = \beta_0 + \beta_t^{treatment} + \epsilon_{ts}$$

Sticker *et al.* 2019, *Robust summarization and inference in proteome-wide label-free quantification*, https://doi.org/10.1101/668863.

And also

- Data independent acquisition (DIA)
- Targets proteomics (SRM, MRM, PRM)
- Post-translational modifications (PTMs)
- Protein-protein interactions
- Sub-cellular localisation
- ...

Laurent Gatto

Computational Biology Group
de Duve Institute, UCLouvain
laurent.gatto@uclouvain.be
https://lgatto.github.io
@lgatto
lgatto
0000-0002-1520-2268
lgatto
Google scholar
Impact story
dissem.in

Acknowledgements Sebastian Gibb and Johannes Rainer (`MSnbase` and *R for Mass Spectrometry*)

Open PhD or post-doc position available at the de Duve Institute, UCLouvain, in Brussels. (For international candidates only).

Slides available at

http://bit.ly/20190725csama

Thank you for your attention