

Mass spectrometry-based proteomics

Laurent Gatto

CSAMA, Brixen, 12 July 2018

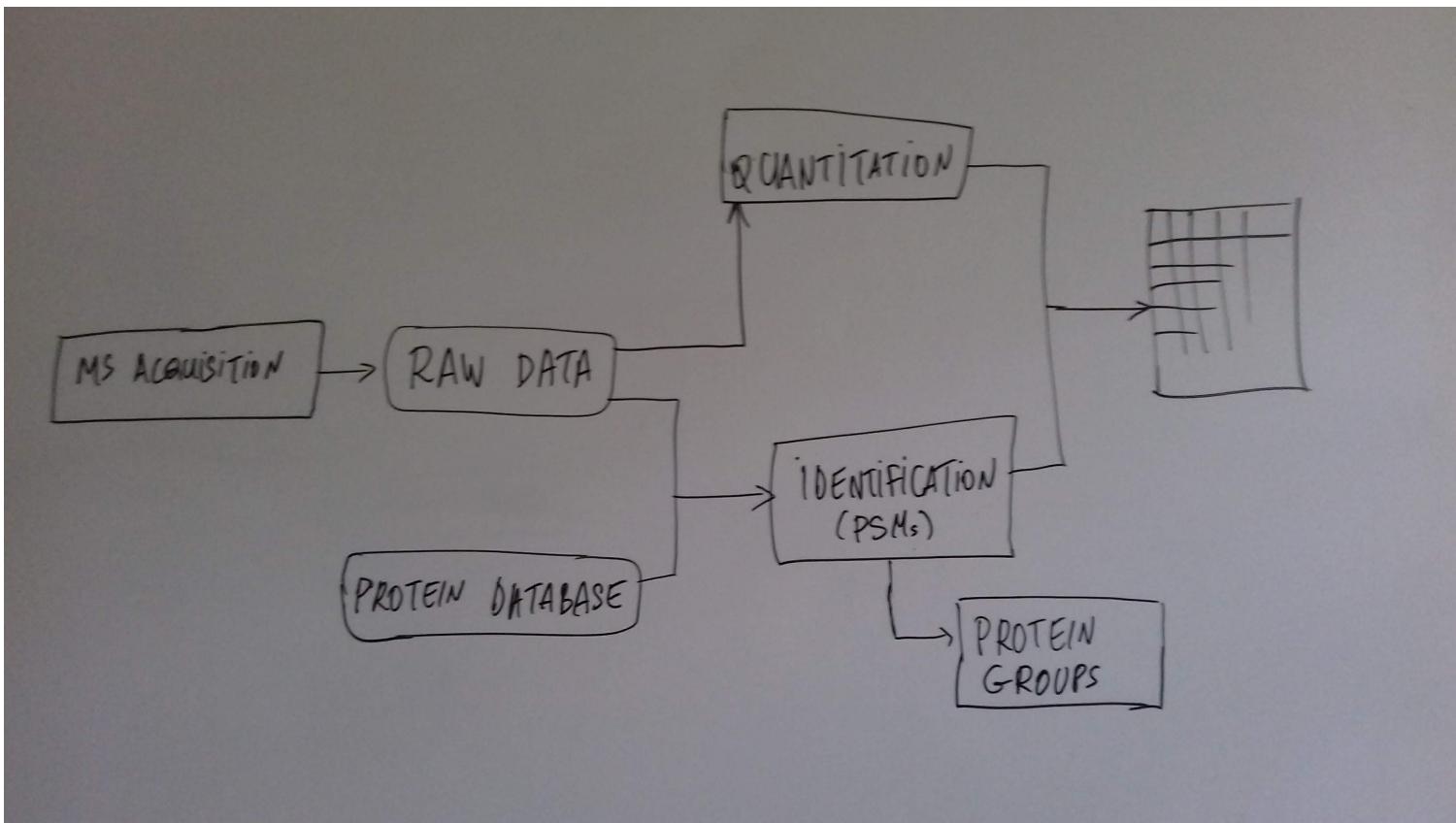
Get the slides at <http://bit.ly/20180712csama>

These slides are available under a **creative common CC-BY license**. You are free to share (copy and redistribute the material in any medium or format) and adapt (remix, transform, and build upon the material) for any purpose, even commercially.

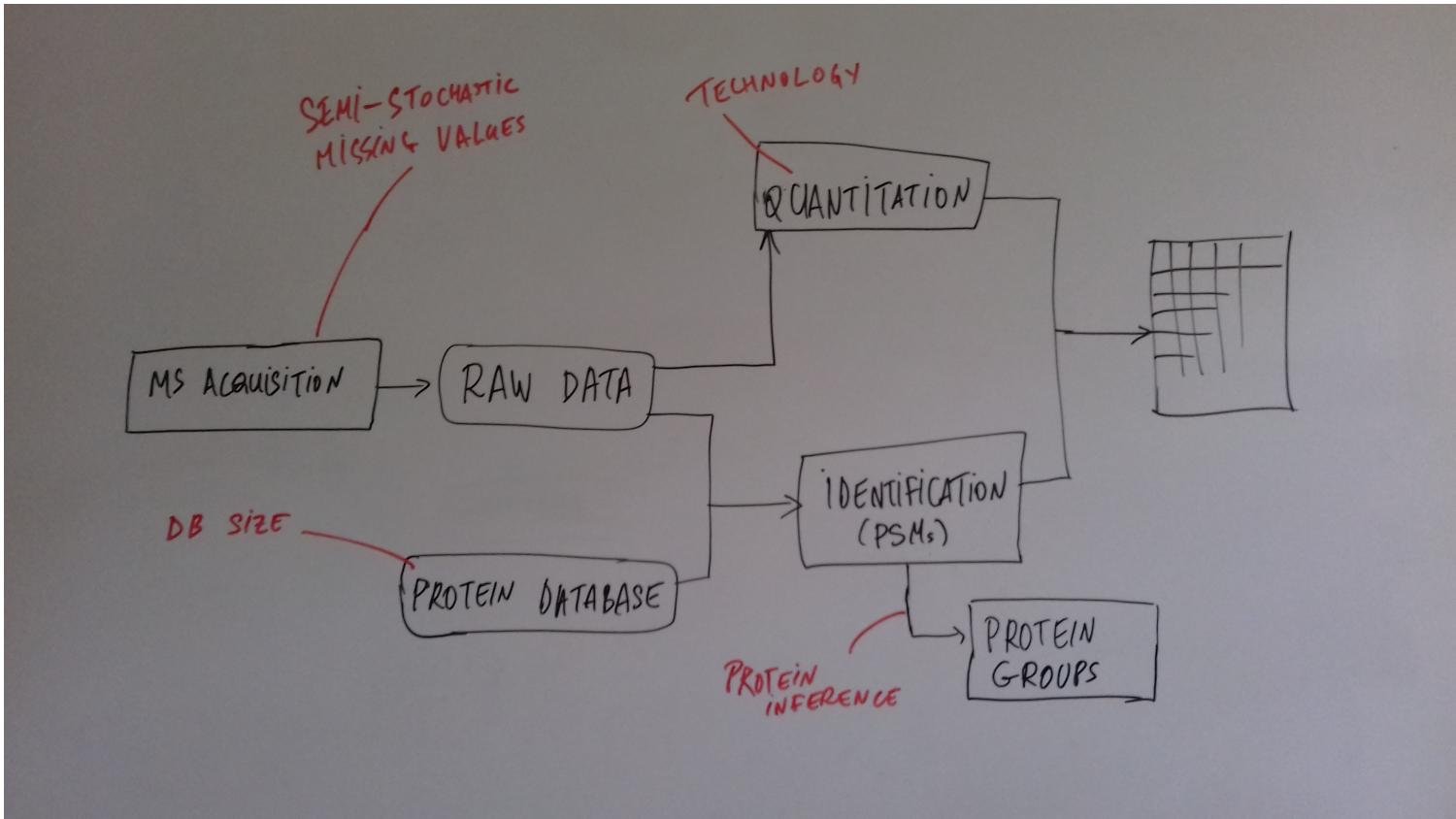


	A	B	C	D	E		A	B	C	D	E
1	ENSG-ID	RPKM U2OS	RPKM U251MG	RPKM A431	RPKM RATIO U2C	1	ENSG-ID	INTENSITY U2O	INTENSITY U251M	INTENSITY A43	SILAC RATIO
2	ENSG00000196433	0	0	0	NA	2	ENSG00000134184	10499000	340820	704250	(
3	ENSG00000166763	0	0	0	NA	3	ENSG00000164828	26281000	9406200	28673000	
4	ENSG00000168781	4.16036	3.877868	2.831423	1.072	4	ENSG00000126945	48935000	24463000	40772000	
5	ENSG00000198746	3.414995	2.182754	2.706358	1.562	5	ENSG00000162722	508090	0	0	NA
6	ENSG00000204131	0.4166259	0	0	NA	6	ENSG00000143653	3840200	7028700	5179900	(
7	ENSG00000134184	0	0	0	NA	7	ENSG00000162851	3545200	3599200	4616700	
8	ENSG00000164828	67.3945	53.05943	144.152	1.270	8	ENSG00000153187	1091800000	606190000	923910000	(
9	ENSG00000126945	9.074128	5.110208	6.99964	1.775	9	ENSG00000203667	13626000	8258900	13803000	
10	ENSG00000185220	1.566899	1.424753	0.9343607	1.099	10	ENSG00000121644	204730	76791	149950	NA
11	ENSG00000171163	4.878578	14.03779	7.697866	0.347	11	ENSG00000035687	66976000	47033000	106990000	
12	ENSG00000171161	7.5040403	15.39958	10.45916	0.487	12	ENSG00000117020	1462300	1002000	1272400	
13	ENSG00000175137	4.635162	14.90939	7.502928	0.310	13	ENSG00000143702	4304700	7686900	3976500	
14	ENSG00000189181	0	0	0	NA	14	ENSG00000203668	4410400	1232100	2063400	:
15	ENSG00000177151	0	0	0	NA	15	ENSG00000091483	147560000	148660000	137460000	(
16	ENSG00000187701	0	0	0	NA	16	ENSG00000116984	2563700	1013400	1934000	
17	ENSG00000184022	0	0	0	NA	17	ENSG00000119285	80282000	39432000	63197000	
18	ENSG00000183130	0	0	0	NA	18	ENSG00000116977	402720	312940	740150	
19	ENSG00000183310	0	0	0	NA	19	ENSG00000143669	49550	40389	153270	
20	ENSG00000182783	0	0	0	NA	20	ENSG00000116957	15326000	11426000	18856000	(
21	ENSG00000188558	0	0	0	NA	21	ENSG00000152904	1257800	982140	1865100	
22	ENSG00000203661	0	0	0	NA	22	ENSG00000188739	4634100	3248500	4772200	
23	ENSG00000196539	0	0	0	NA	23	ENSG00000173726	4458400	3902200	5744800	
24	ENSG00000196240	0	0	0	NA	24	ENSG00000168264	885950	797480	1048700	
25	ENSG00000198104	0	0	0	NA	25	ENSG00000168275	1024600	1675900	1146900	
26	ENSG00000175143	0	0	0	NA	26	ENSG00000135778	11771000	3641700	4445400	:
27	ENSG00000196944	0	0	0	NA	27	ENSG00000116918	25919000	10251000	13153000	(
28	ENSG00000177174	0	0	0	NA	28	ENSG00000135766	252780	296600	394150	
29	ENSG00000177201	0	0	0	NA	29	ENSG00000116903	2575100	1273000	2282500	(
30	ENSG00000177186	0	0	0	NA	30	ENSG00000119280	1411300	150880	521110	:
31	ENSG00000177212	0	0	0	NA	31	ENSG00000099977	172740000	49442000	180810000	

Overview



Overview



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.

How does MS work?

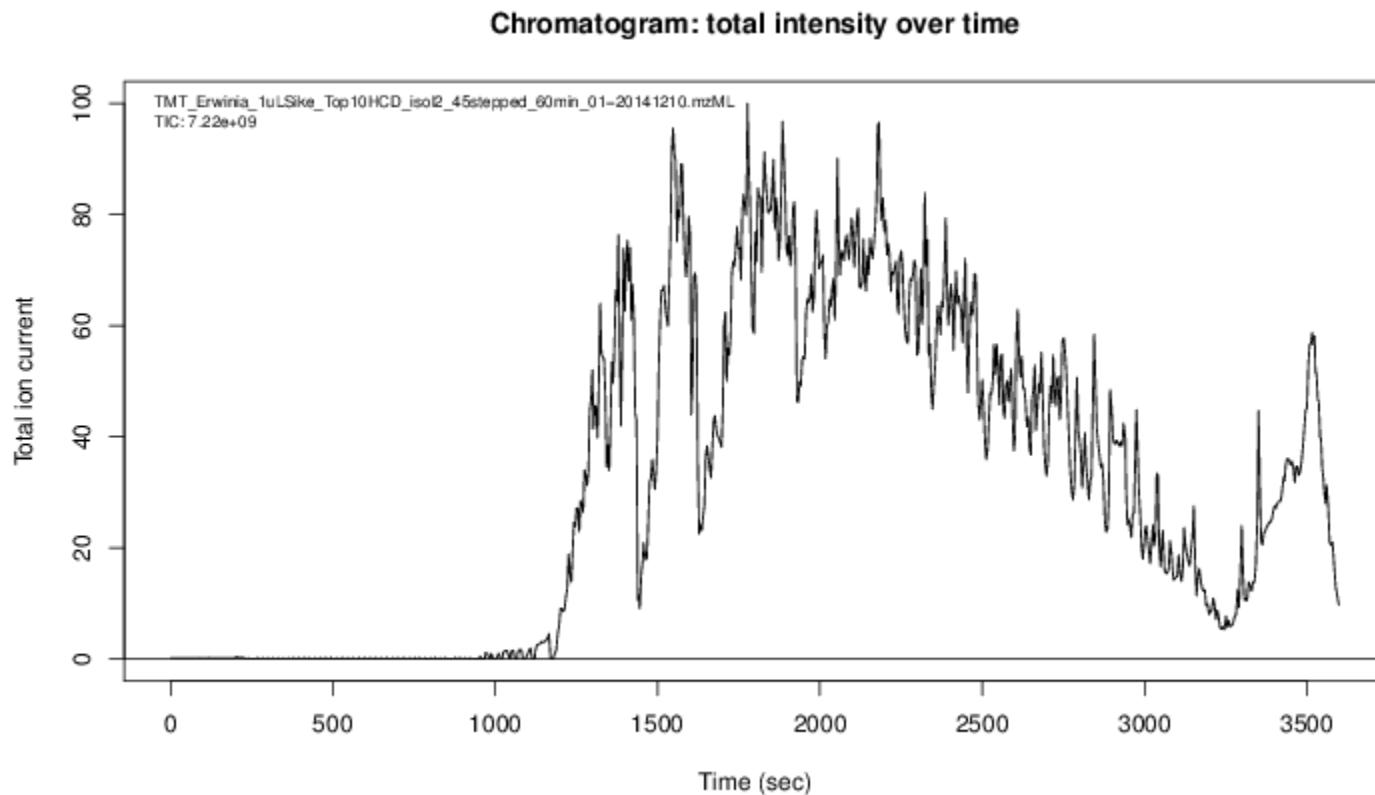
1. Digestion of proteins into peptides - as will become clear later, the features we measure in shotgun (or bottom-up) *proteomics* are peptides, **not** proteins.
2. On-line liquid chromatography (LC-MS)

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

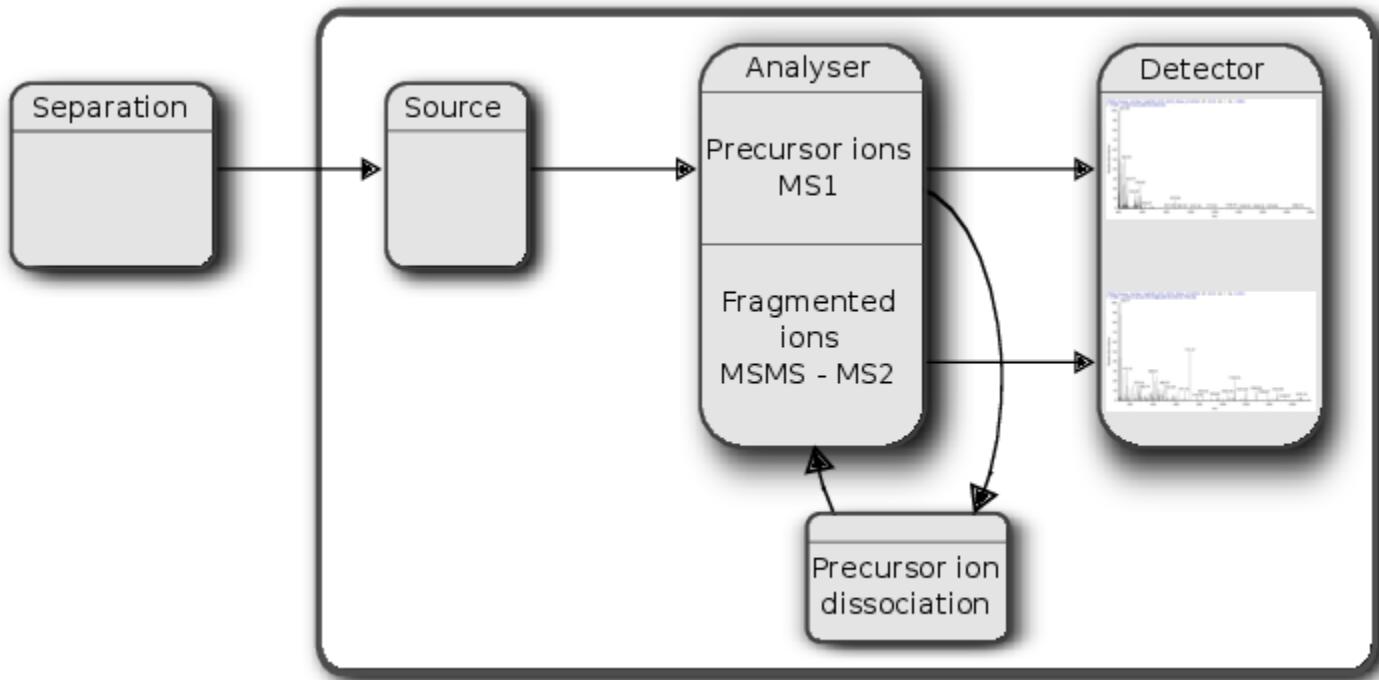
MS is generally coupled to chromatography (liquid LC, but can also be gas-based GC). The time an analytes takes to elute from the chromatography column is the **retention time**.



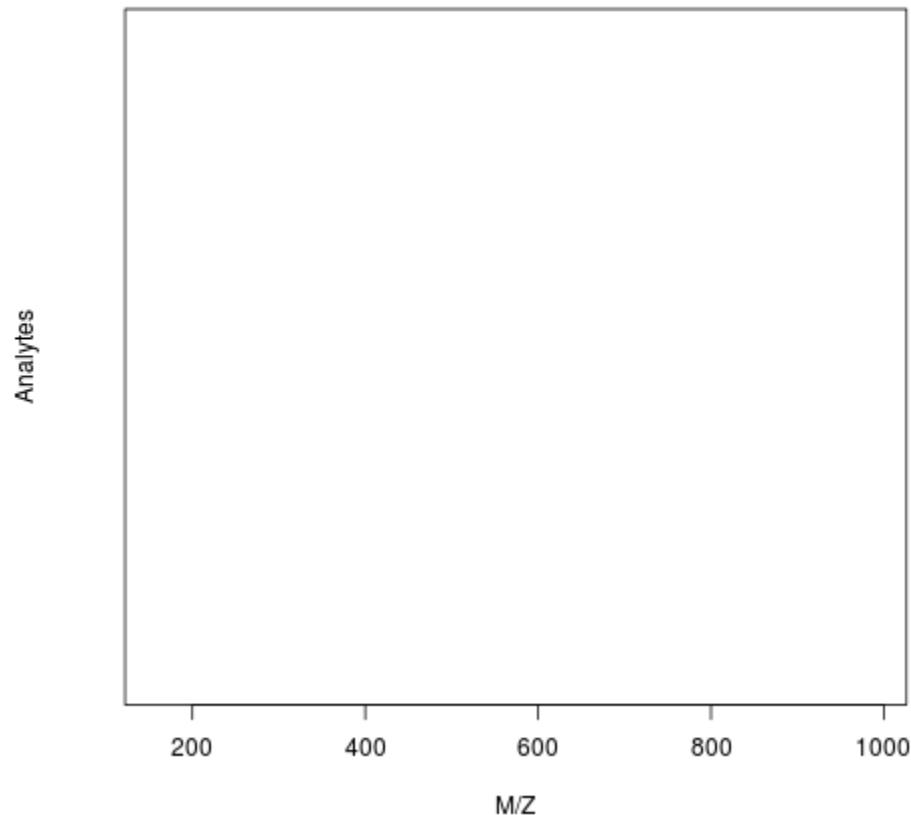
An mass spectrometer is composed of three components:

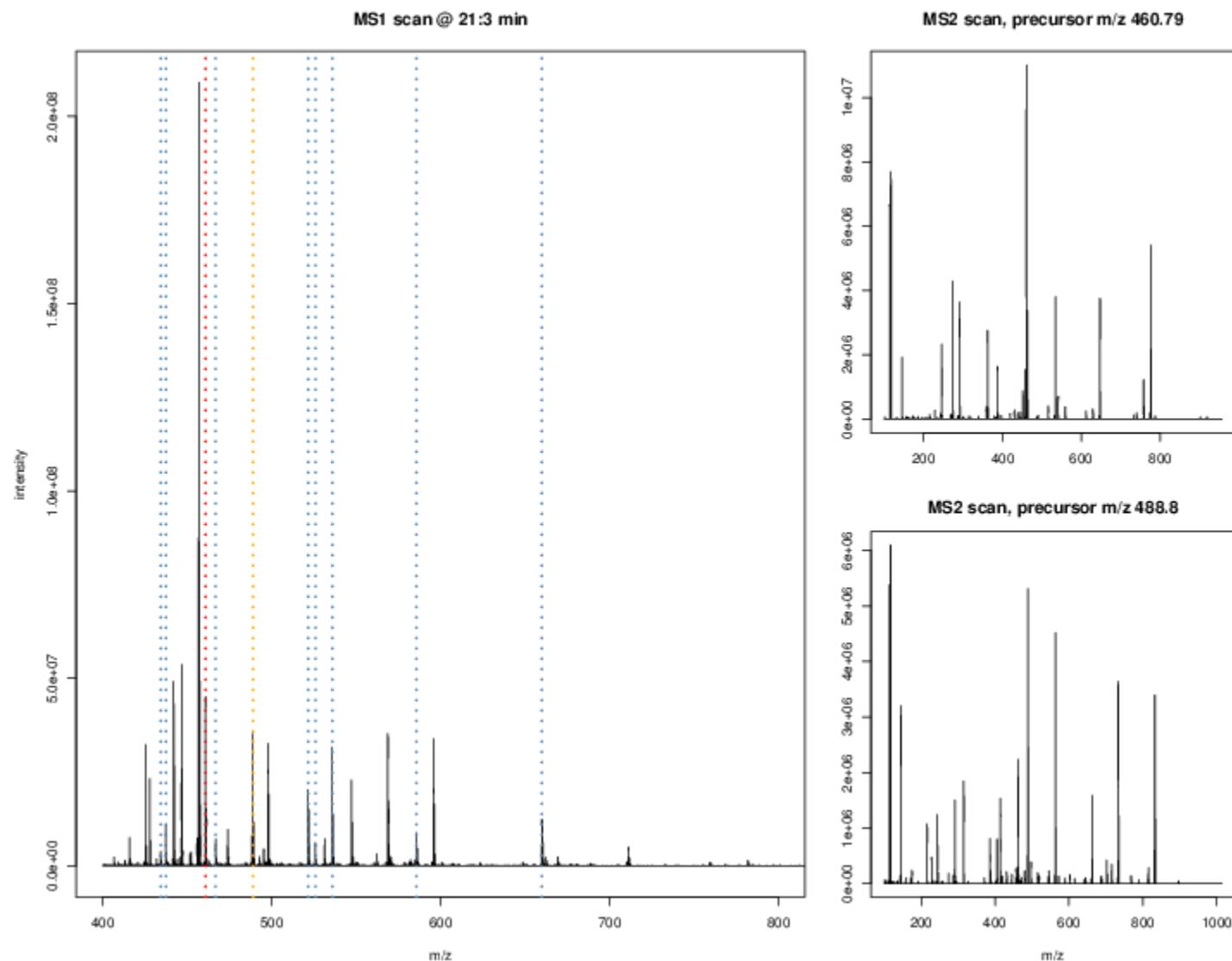
1. The *source*, that ionises the molecules: examples are Matrix-assisted laser desorption/ionisation (MALDI) or electrospray ionisation (ESI).
2. The *analyser*, that separates the ions: Time of flight (TOF) or Orbitrap.
3. The *detector* that quantifies the ions.

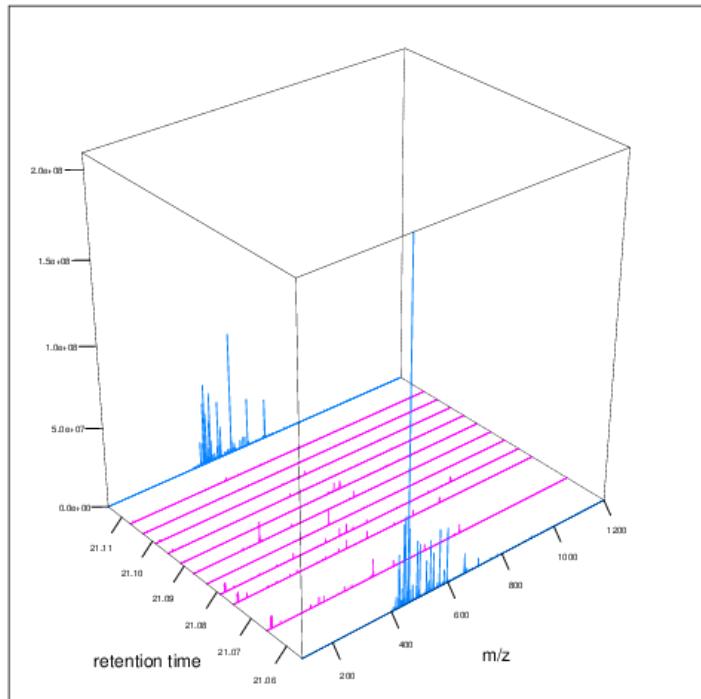
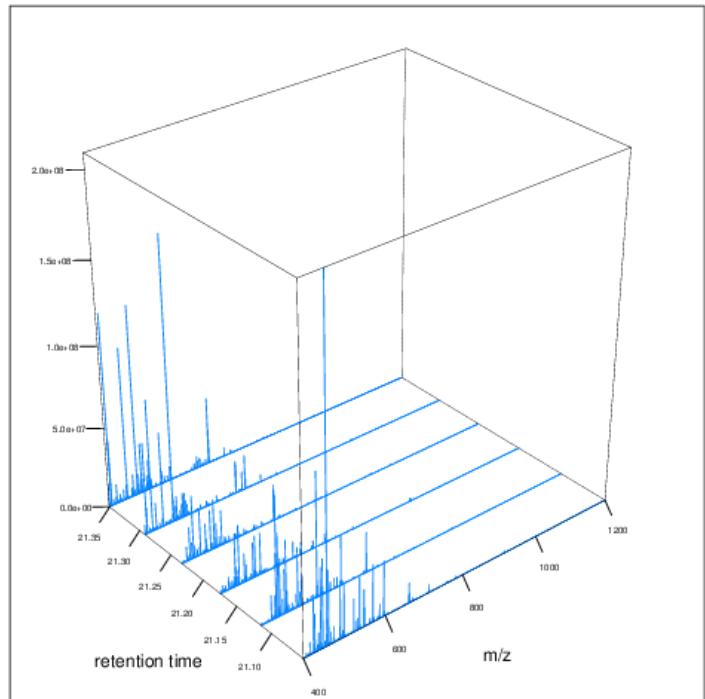
Ions typically go through that cycle at least twice (MS2, tandem MS, or MSMS). Before the second cycle, individual *precursor* ions are selected and broken into *fragment* ions.



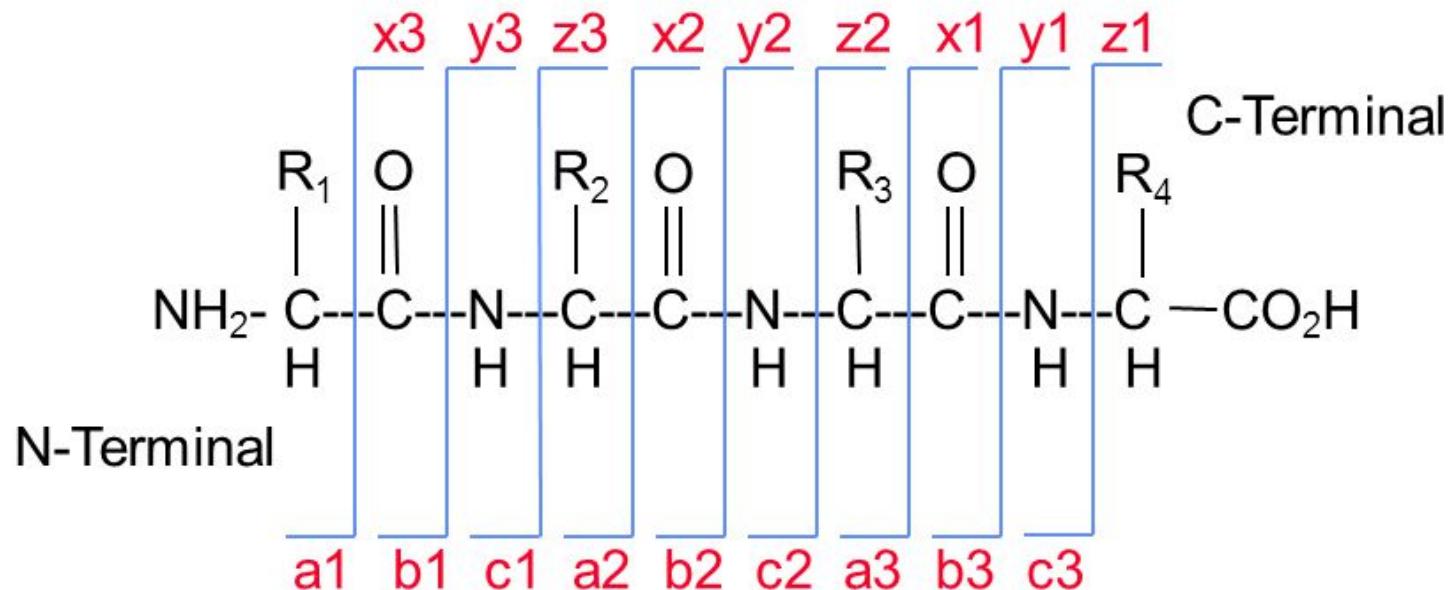
Analyser (1/10)







Identification: fragment ions



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

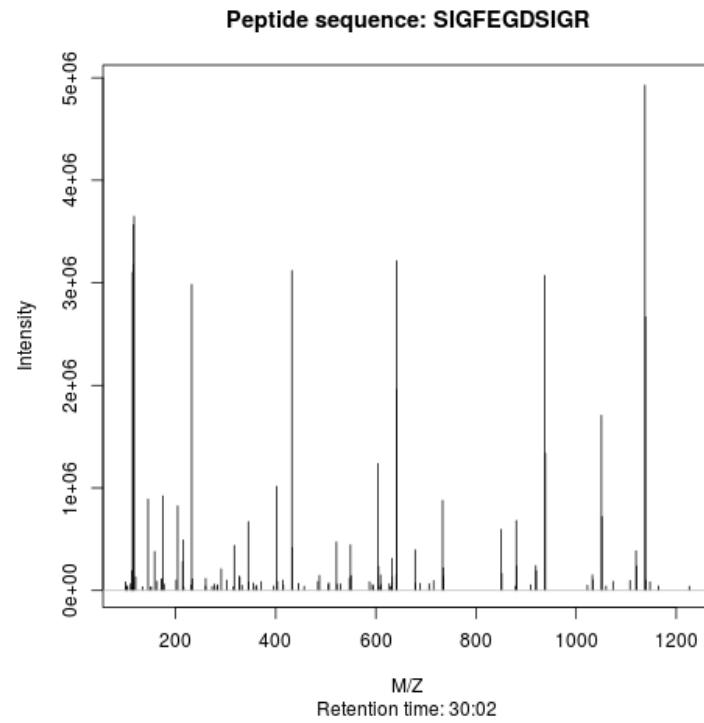
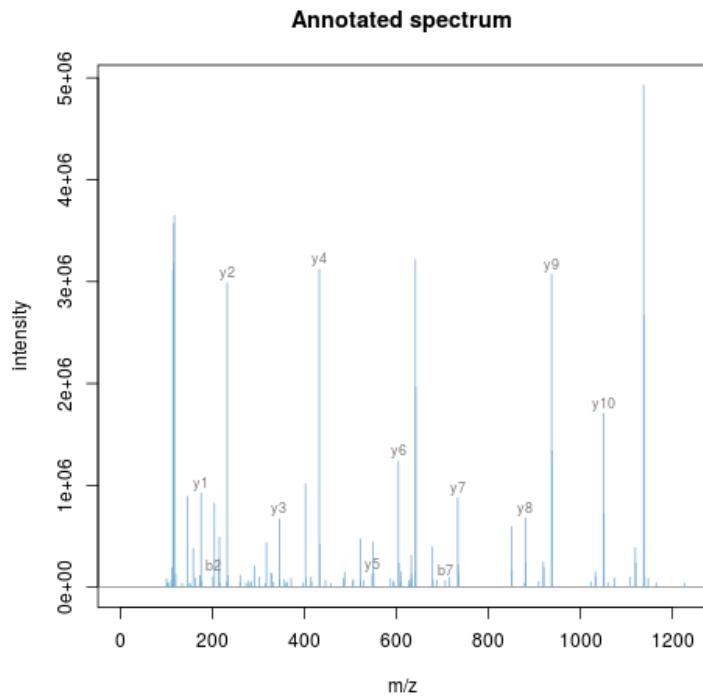
Identification: Peptide-spectrum matching (PSM)

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

 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.

© news.nationalgeographic.com

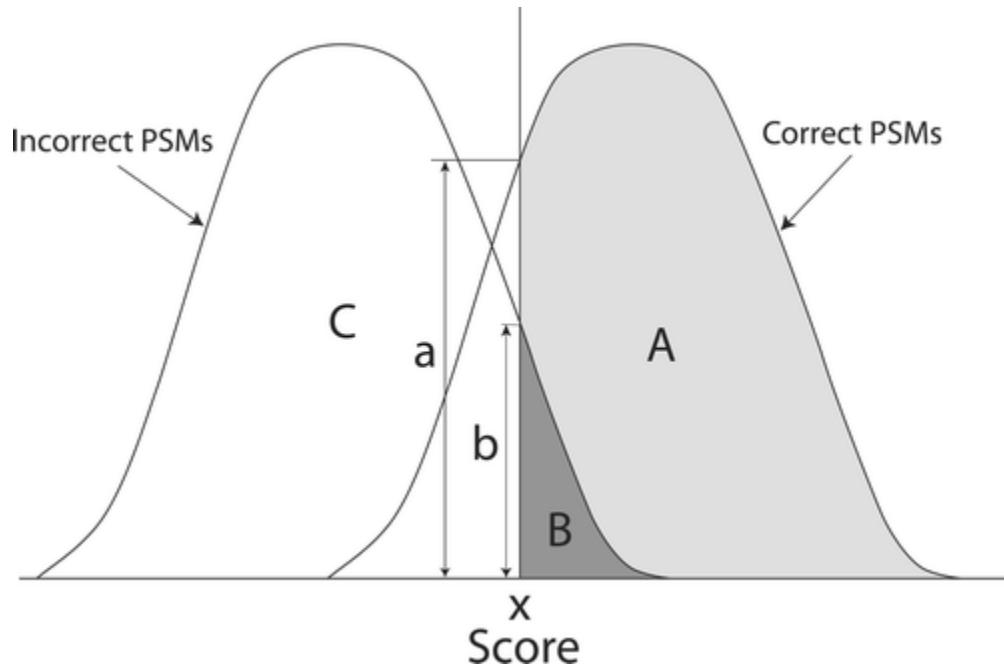
Overall life expectancy in Europe is 81 years.

Componentsⁱ

[Download](#) View all proteins

<input type="checkbox"/>	Component name	Genome Accession(s)	<input type="checkbox"/> Proteins
<input type="checkbox"/>	Chromosome 1	CM000663	5563
<input type="checkbox"/>	Chromosome 2	CM000664	4596
<input type="checkbox"/>	Chromosome 3	CM000665	4122

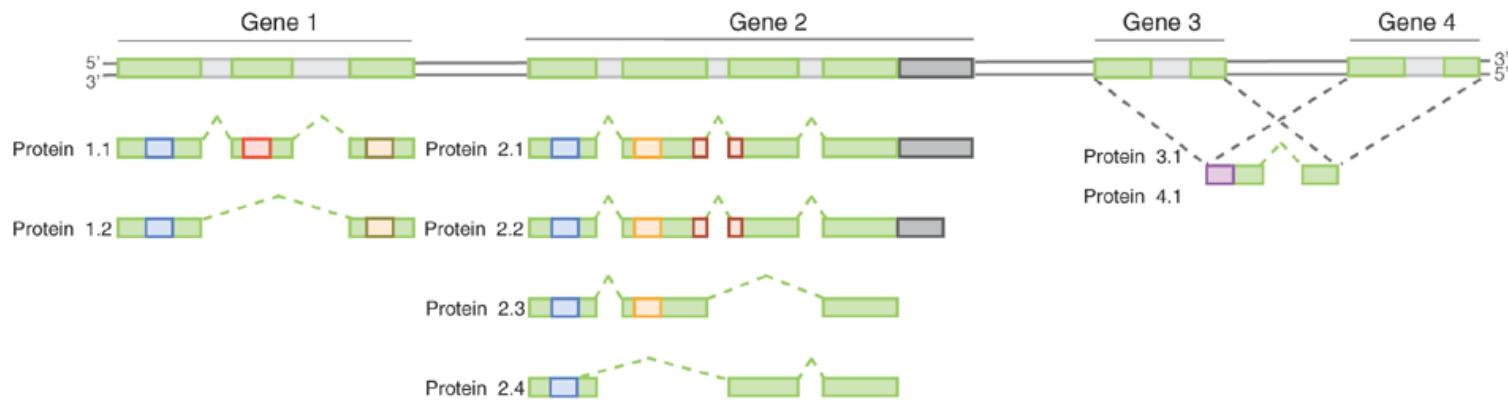
Identification



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

Identification: Protein inference

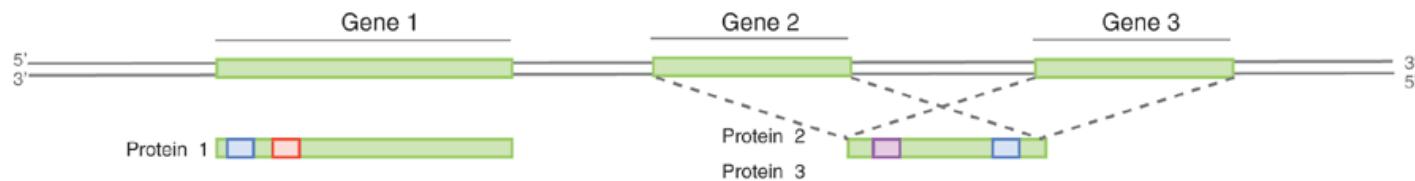
- Keep only reliable peptides
- From these peptides, infer proteins
- If proteins can't be resolved due to shared peptides, merge them into **protein groups** of indistinguishable or non-differentiable proteins.



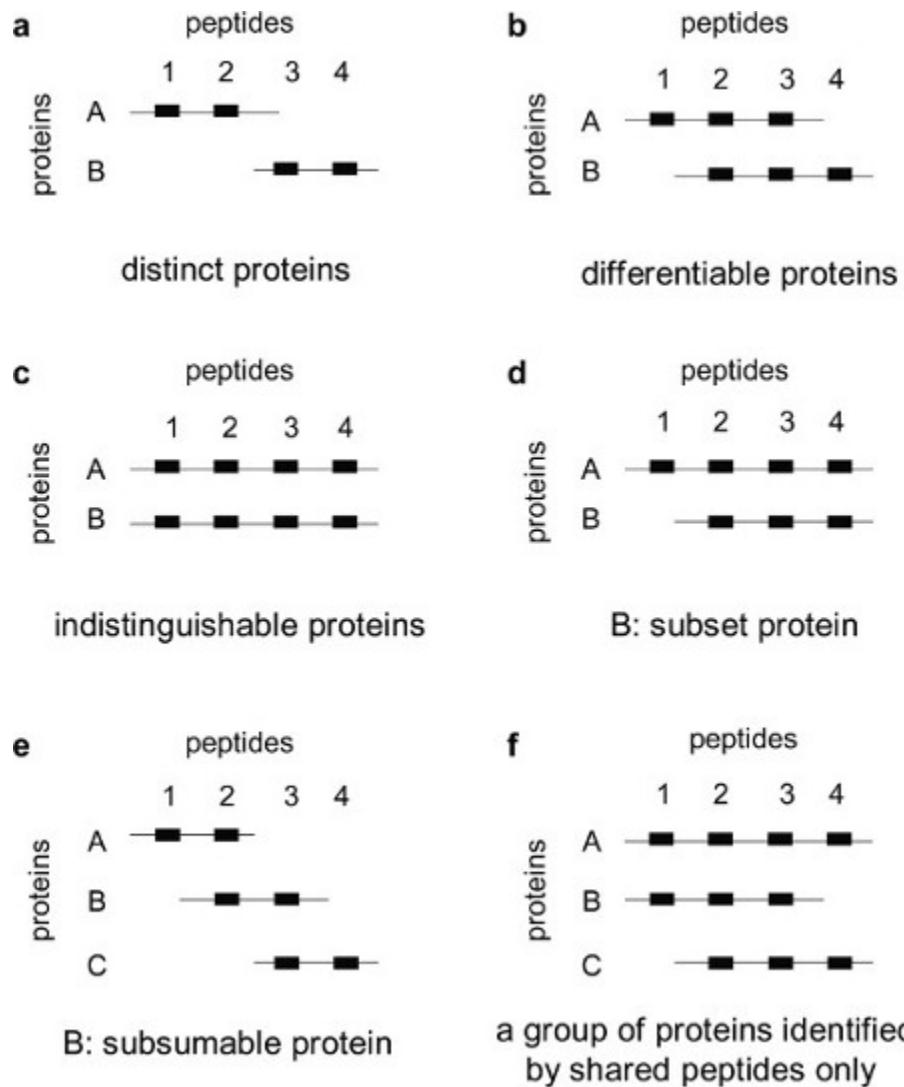
Eukaryotes

Prokaryotes

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).

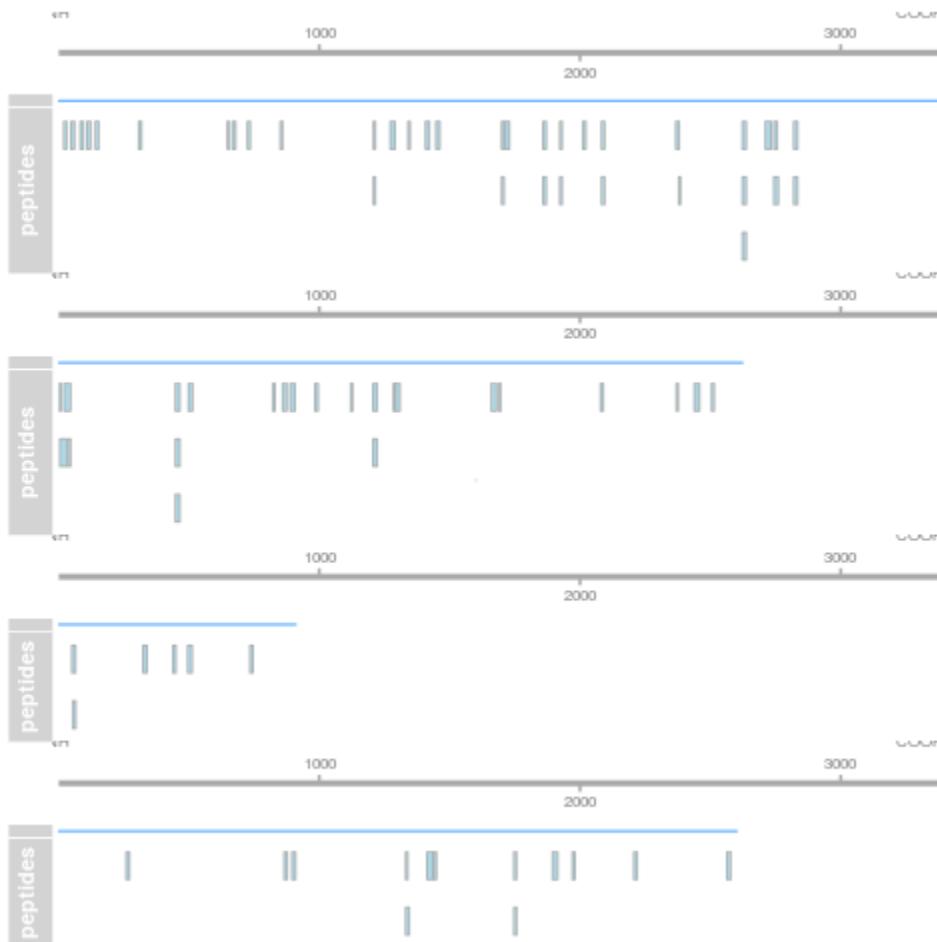


From Nesvizhskii and Aebersold (2005).

Quantitation

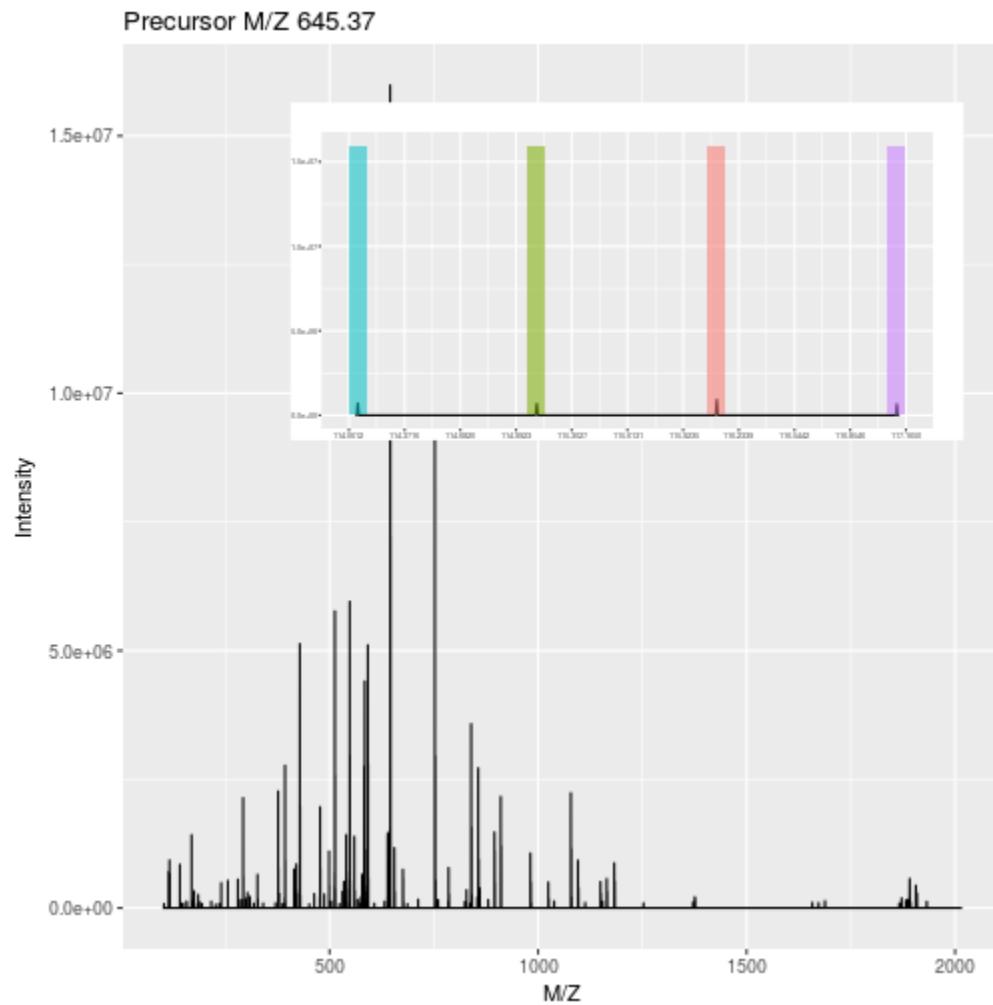
	Label-free	Labelled
MS1	XIC	SILAC, 15N
MS2	Counting	iTRAQ, TMT

Label-free MS2: Spectral counting

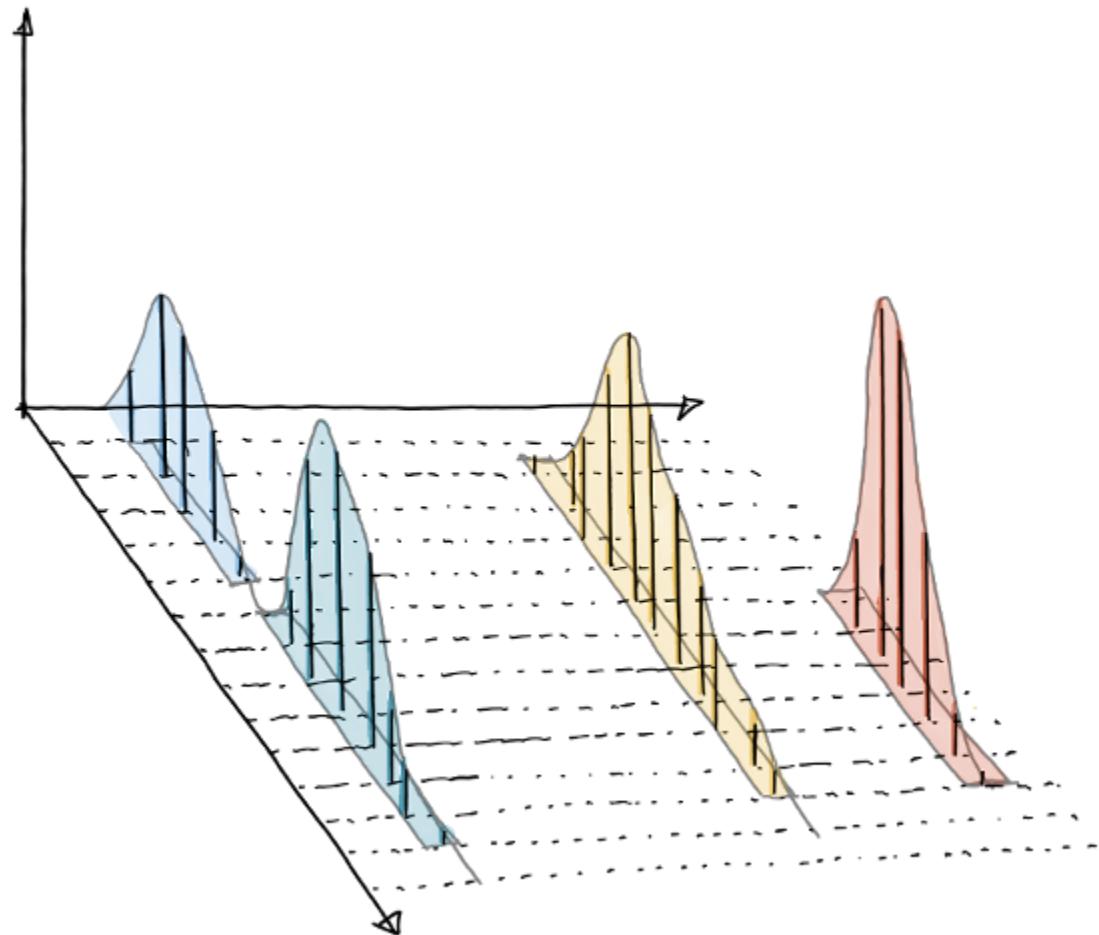


From the [Pbase](#) package.

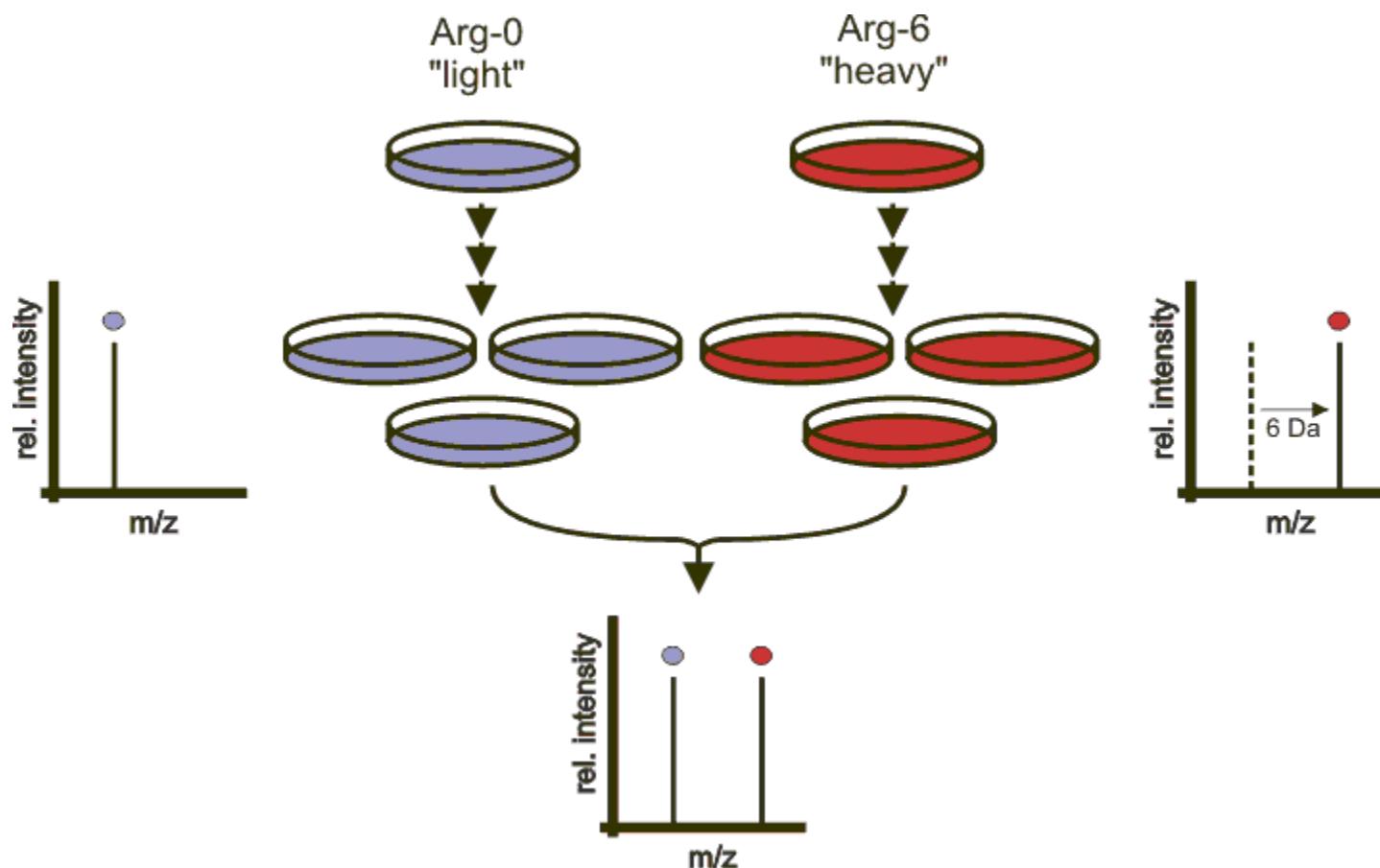
Labelled MS2: Isobaric tagging



Label-free MS1: extracted ion chromatograms



Labelled MS1: SILAC



Credit: Wikimedia Commons.

Quantitation

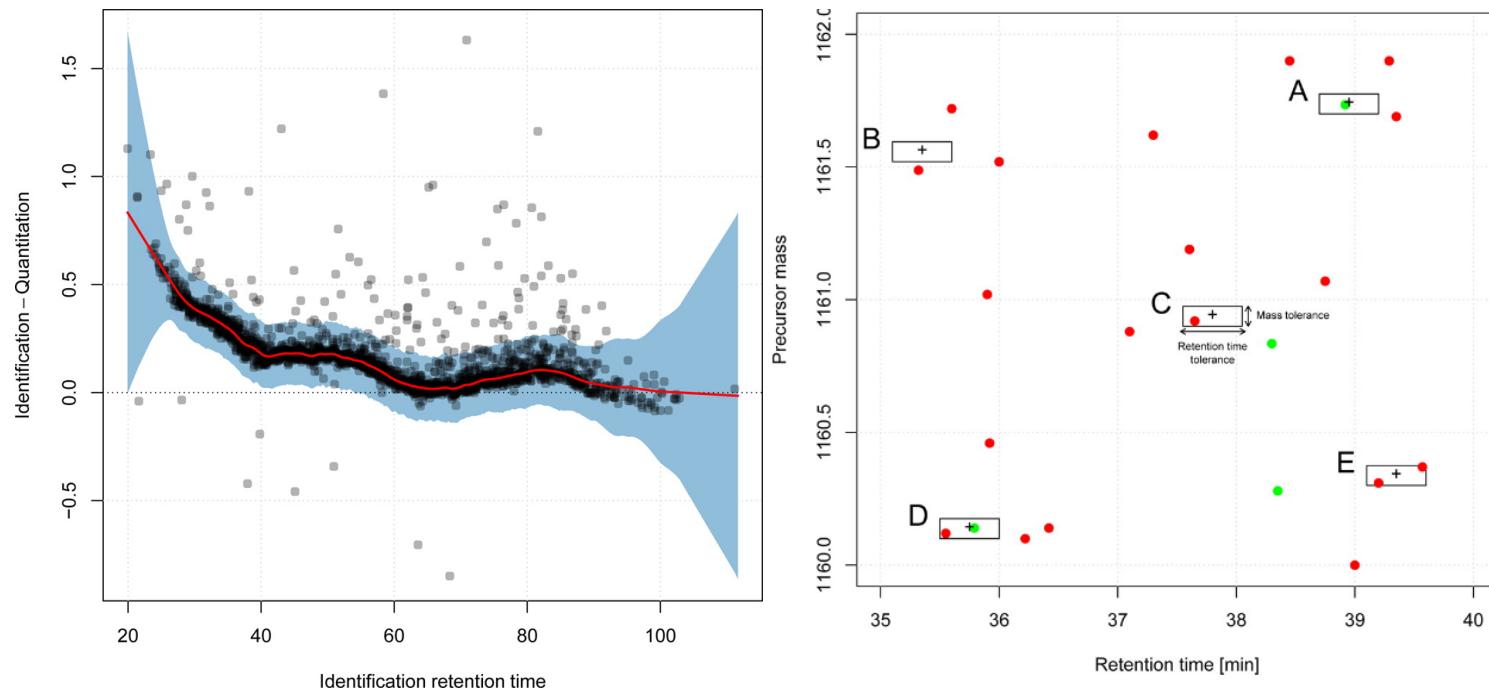
	Label-free	Labelled
MS1	XIC	SILAC, 15N
MS2	Counting	iTRAQ, TMT

- Labelled methods allow to pool samples before MS acquisition (up to 10 samples for Tandem Mass Tags (TMT)), rather than running individual MS runs for each samples, and thus sharing technical variability and reducing missing data (see for example Russel and Lilley [Pipeline to assess the greatest source of technical variance in quantitative proteomics using metabolic labelling](#)).
- Summarising spectrum/peptide intensities into the protein (group) level quantities.

Dealing with missing values

Really missing, or not detected (semi-stochastic nature of precursor selection in shotgun proteomics)?

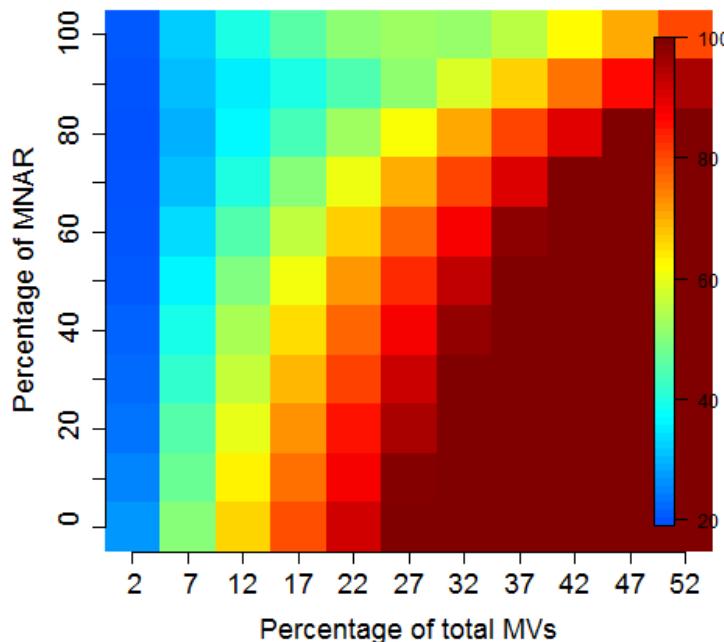
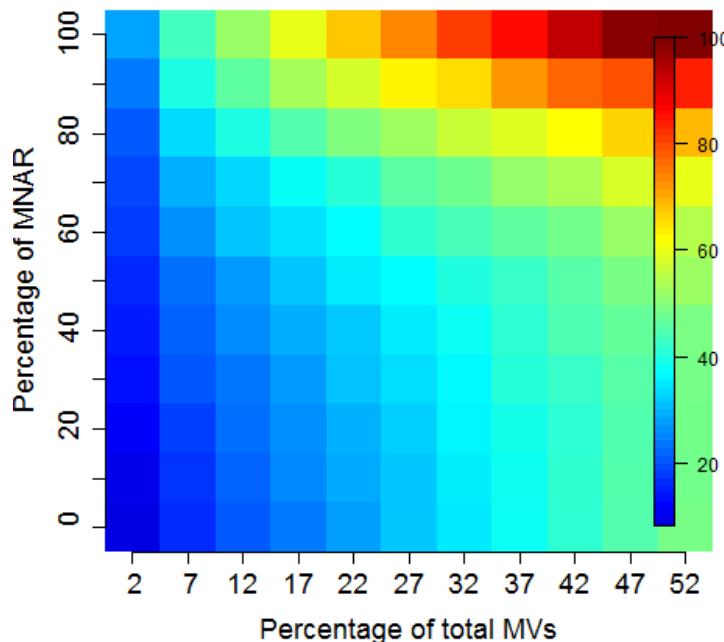
- Identification transfer/matching between runs.
- Filtering.
- Imputation.



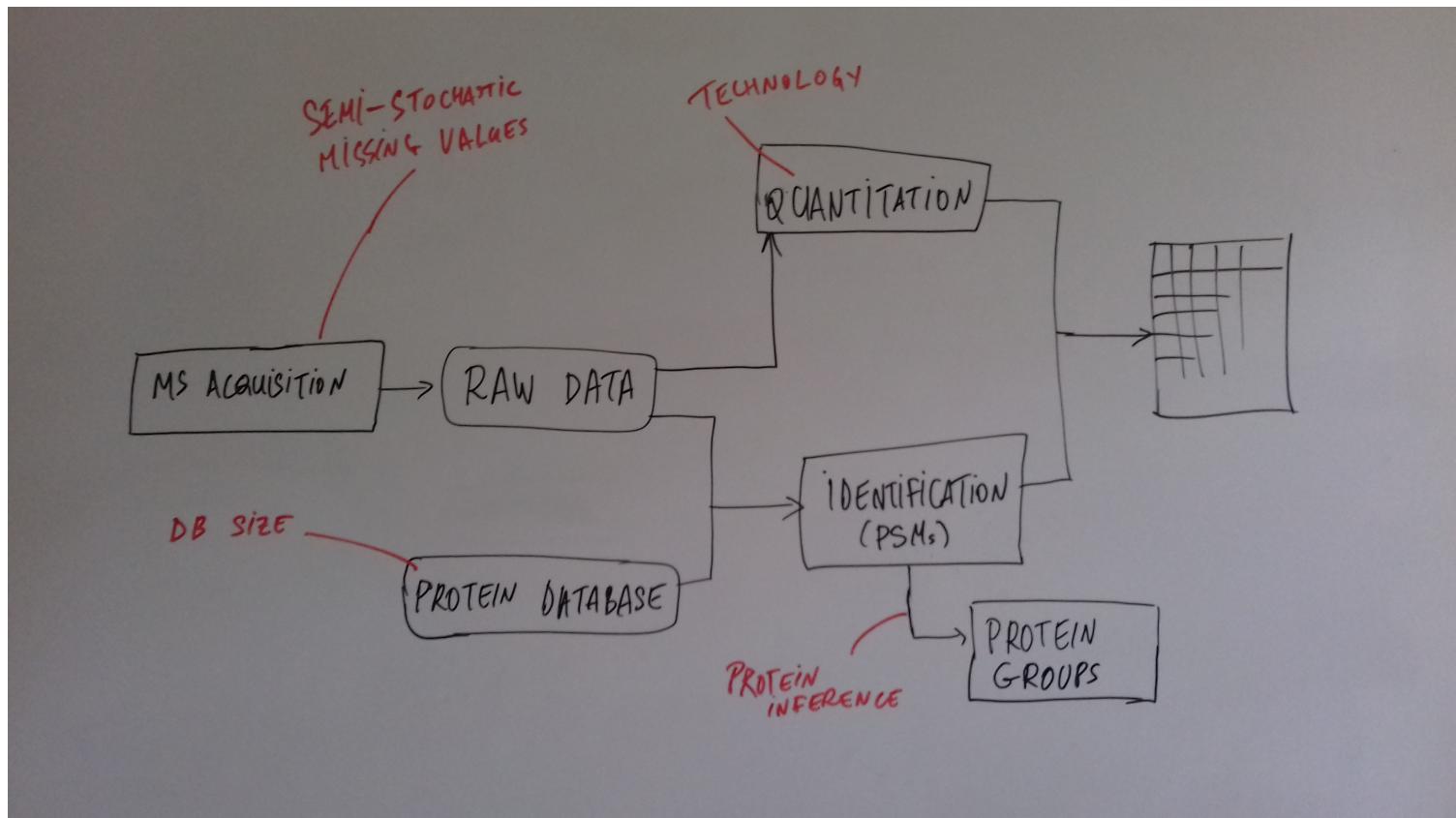
From Bond *et al.* Improving Qualitative and Quantitative Performance for MSE-based Label-free Proteomics.

Missing values

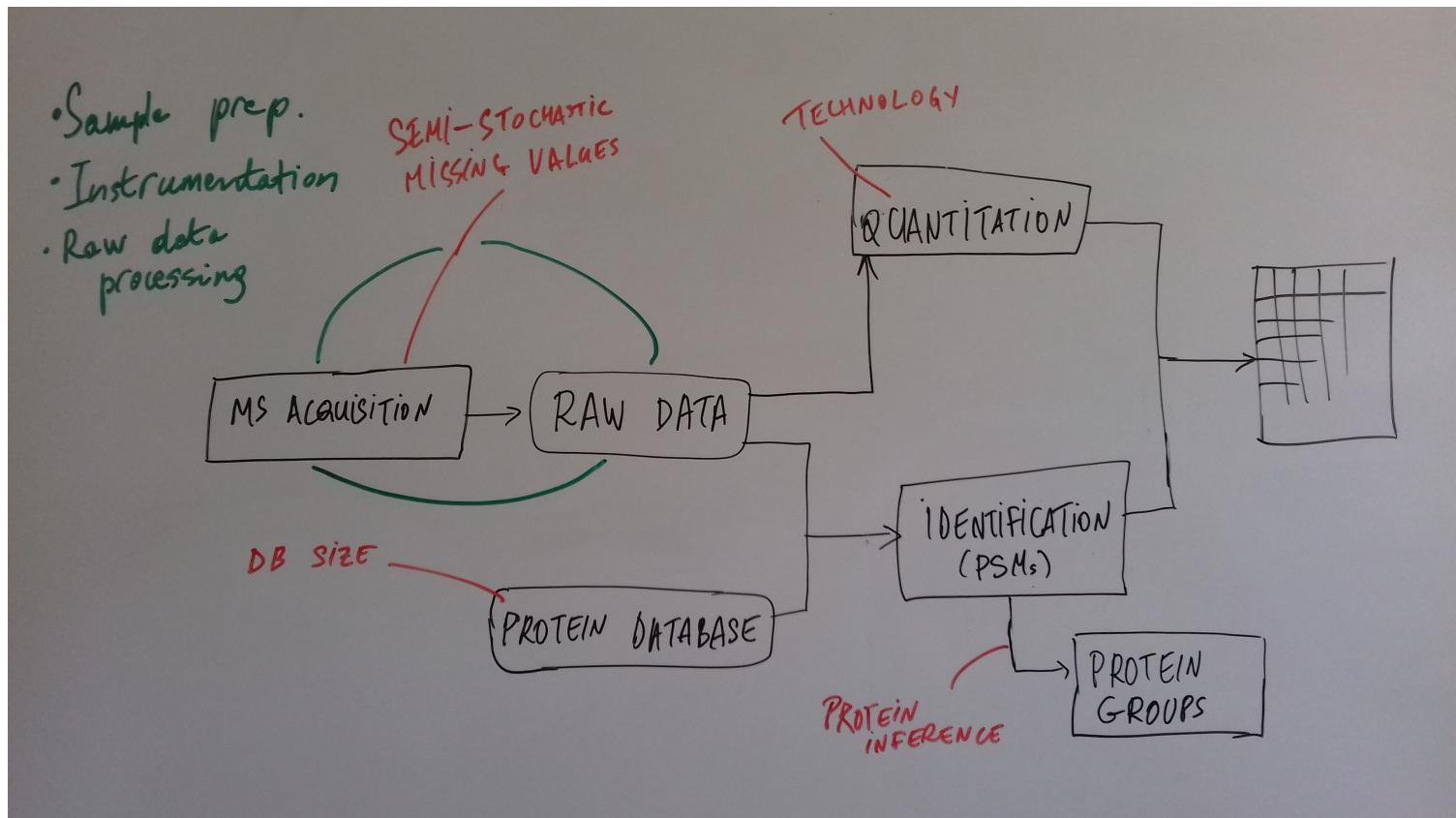
- Filtering.
- Imputation: considering the underlying nature of missingness, i.e missing not at random (left-censored) or at random.



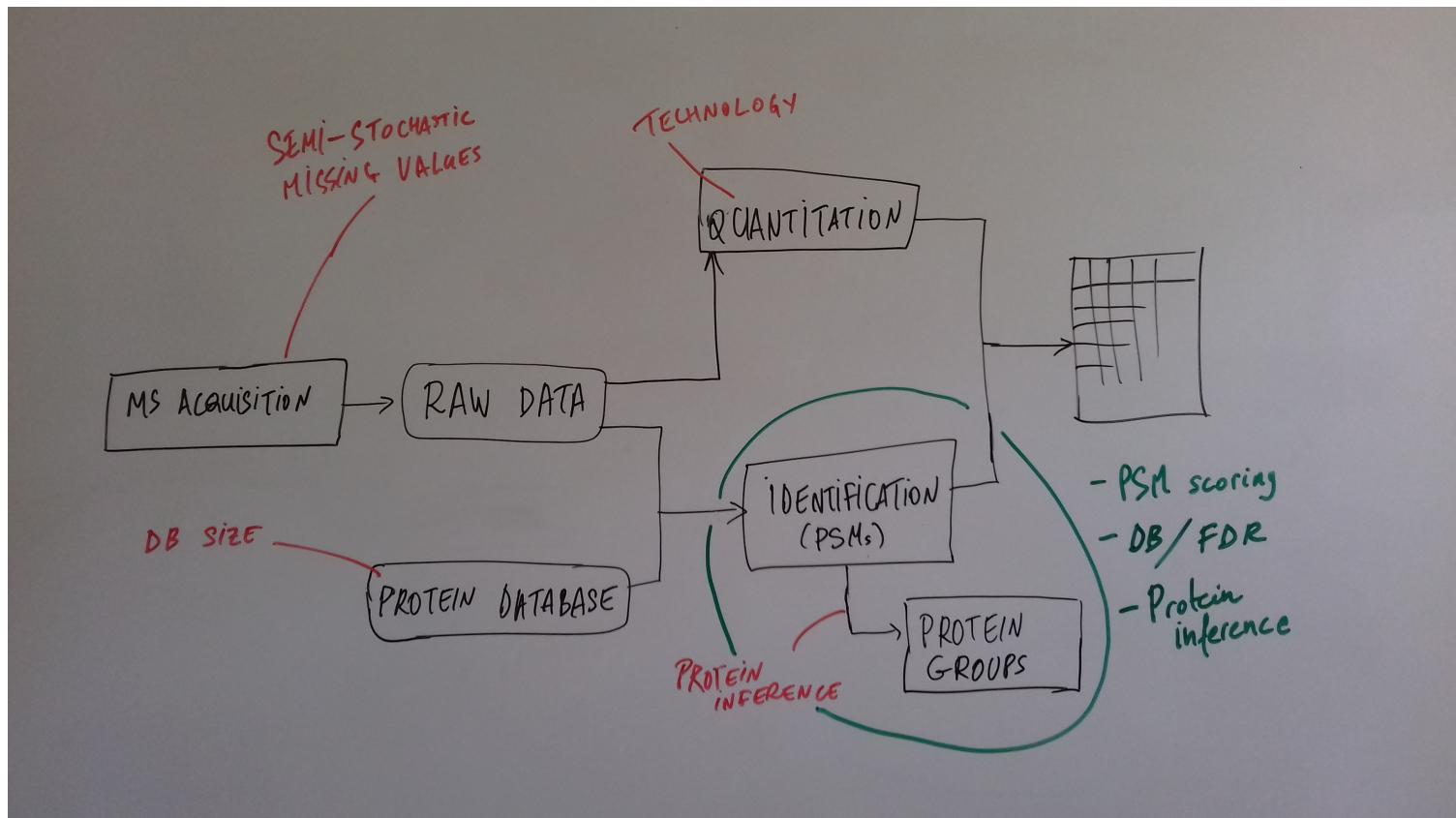
Recap



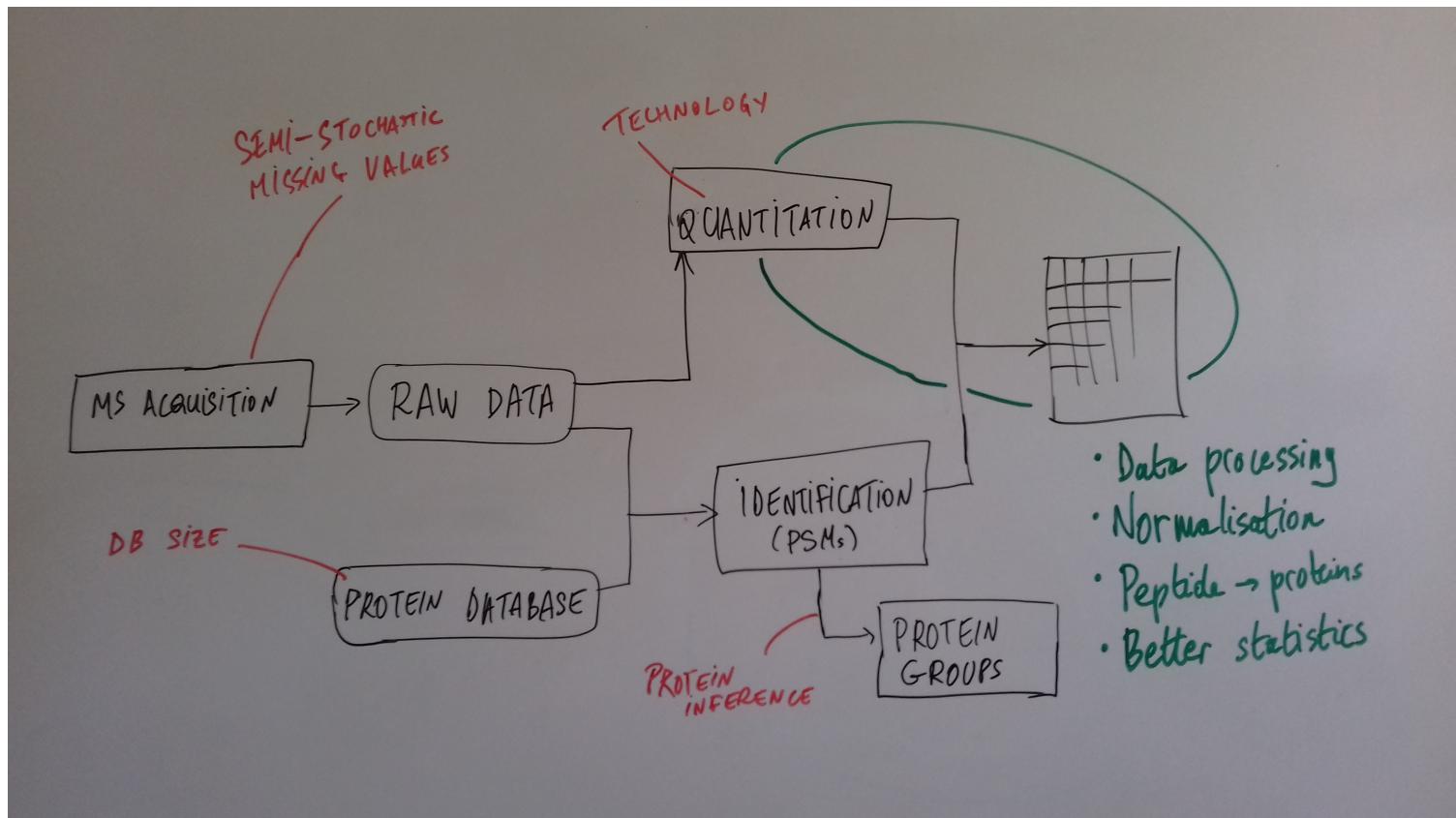
Challenges



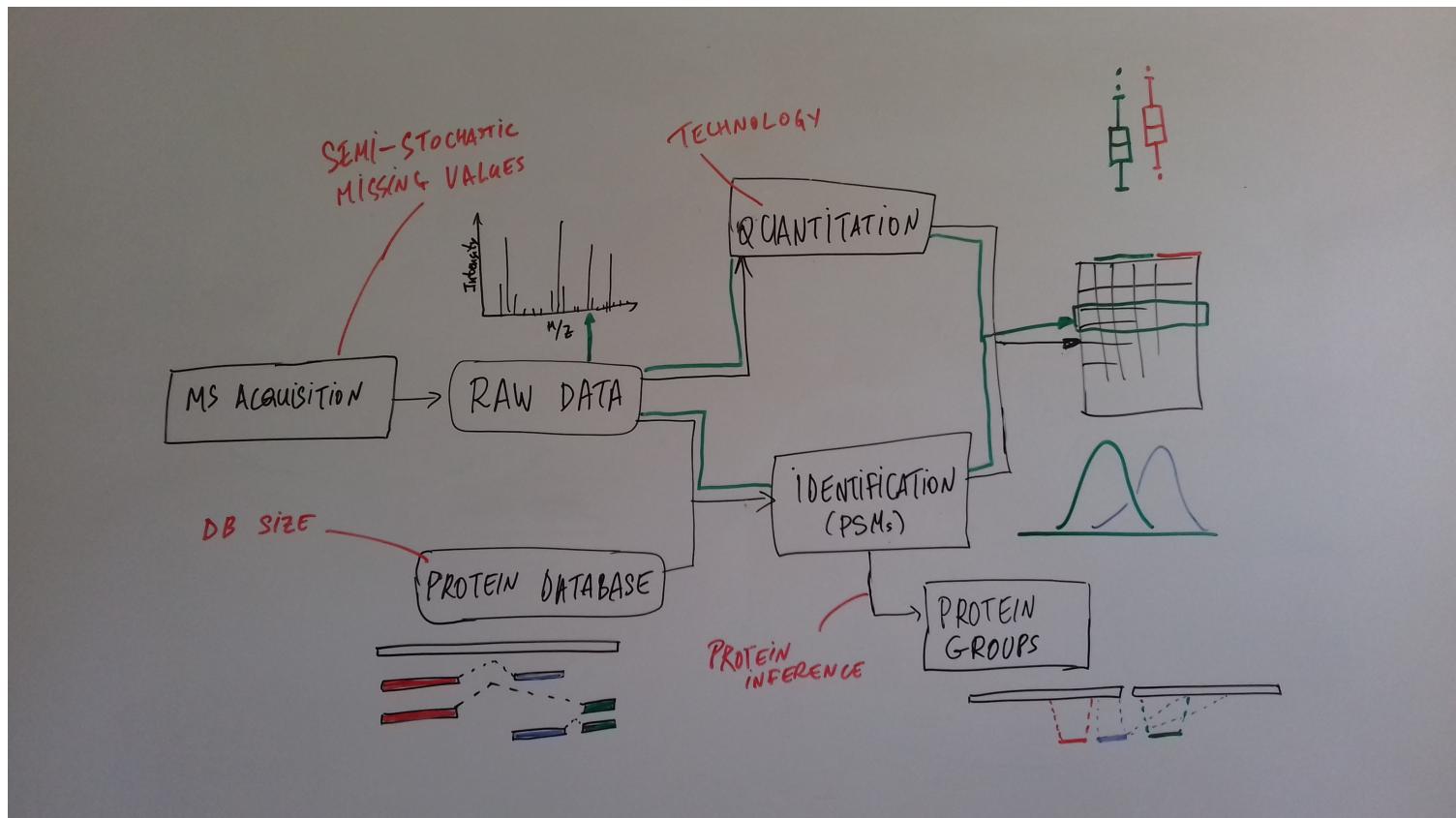
Challenges



Challenges



Challenges



Bioconductor software infrastructure

Raw data: mzML

`MSnbase::readMSData` to read raw data into R as `MSnExp`.

Identification data: mzid

`MSnbase::readMzIdData` as a `data.frame`.

Quantitative data

`MSnbase::readMSData2` as an `MSnSet`.

Other technologies: DIA

Data Independent Acquisition (DIA) (as opposed to Data Dependent Acquisition - DDA).

Examples:

- MS^E , where the whole M/Z range is fragmented (at different energies) and precursors and fragment ions are matched using their elution profiles, by Waters.
- SWATH-MS (Sequential Windowed Acquisition of All Theoretical Fragment Ion Mass Spectra) by AB Sciex
- AIF (All-ion fragmentation, similar to MS^E) by Thermo Fisher.
- ...

The advantage of these DIA methods is that all the data are recorded; they have the potential to analyse all the data, thus limiting the number of missing values.

See Law and Lim [Recent advances in mass spectrometry: data independent analysis and hyper reaction monitoring](#).

Other technologies: targeted proteomics

Selected/multiple/parallel reaction monitoring (SRMs, MRM, PRM) are pre-defined MS features that the instrument will specifically scan for. This is the method to assay a limited number of peptides/proteins in a high number of samples.

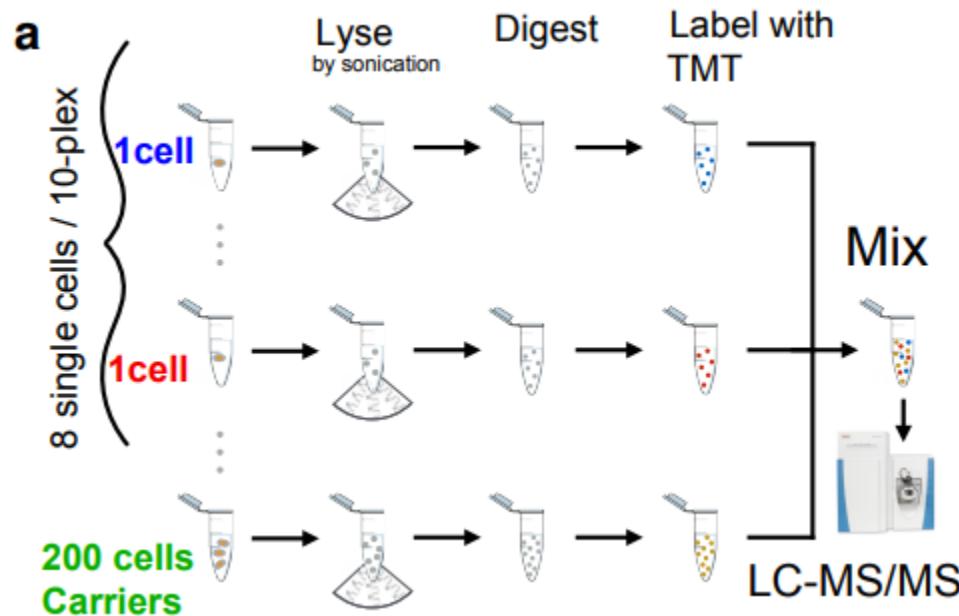
Other technologies: single cell proteomics using CyTOF

In **mass cytometry**, rare earth metals are used as reporters on antibodies. Analysis of metal abundances using the mass cytometer allows determination of the expression of 100s of markers in 1000s individual cells.

See Giesen *et al.* [Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry](#) and Bendall *et al.* [Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum](#).

Other technologies: single cell proteomics using MS

Single Cell ProtEomics by Mass Spectrometry (SCoPE-MS) can quantify thousands of proteins in single mammalian cells.



Budnik *et al.* Mass-spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation.

See also Zhu *et al.* Nanodroplet processing platform for deep and quantitative proteome profiling of 10–100 mammalian cells.

Additional dimensions

- Post-translational modifications (PTMs)
- Protein-protein interactions (PPI)
- Protein sub-cellular localisation (spatial proteomics)

When based on MS, all points discussed above will remain relevant.



Laurent Gatto

📍 University of Cambridge
📍 [de Duve Institute](#), UCLouvain

🏡 <https://lgatto.github.io>
🐦 @lgatto
💻 lgatto
>ID 0000-0002-1520-2268
👤 lgatto
 גוגל Google scholar
👤 Impact story
📝 dissemin.in

Acknowledgements: BBSRC and the Wellcome Trust and collaborators/contributors [Lisa Breckels](#), [Olly Crook](#), [Sebastian Gibb](#) and [Johannes Rainer](#).

New positions available at the de Duve Institute, in Brussels.

Thank you for your attention

Find the slides at
<http://bit.ly/20180712csama>