Proteomics

Jiri Petrak
BIOCEV
First Faculty of Medicine
Charles University, Vestec

Clinical Proteomics Group

www.petraklab.cz

BIOCEV

Biotechnology and Biomedicine Centre of the Academy
of Sciences and Charles University in Vestec

Introduction to proteomics

- Human proteome
- Proteomic workflow
- Mass spectrometry and protein identification

An example of a proteomic study

- A quest for circultaing biomarkers of advanced heart failure

„Membrane Wednesday“ – Integral membrane proteins.
Special strategies and tools.
Tumor plasma mebrane proteomics

Proteome – a complete set of proteins present at the given moment in a studied organism, tissue or cell. It includes protein PTM, localization, turnover and protein interactions.

Proteomics – a set of methods and approaches developed and used for quantitative and qualitative characterization of proteomes.

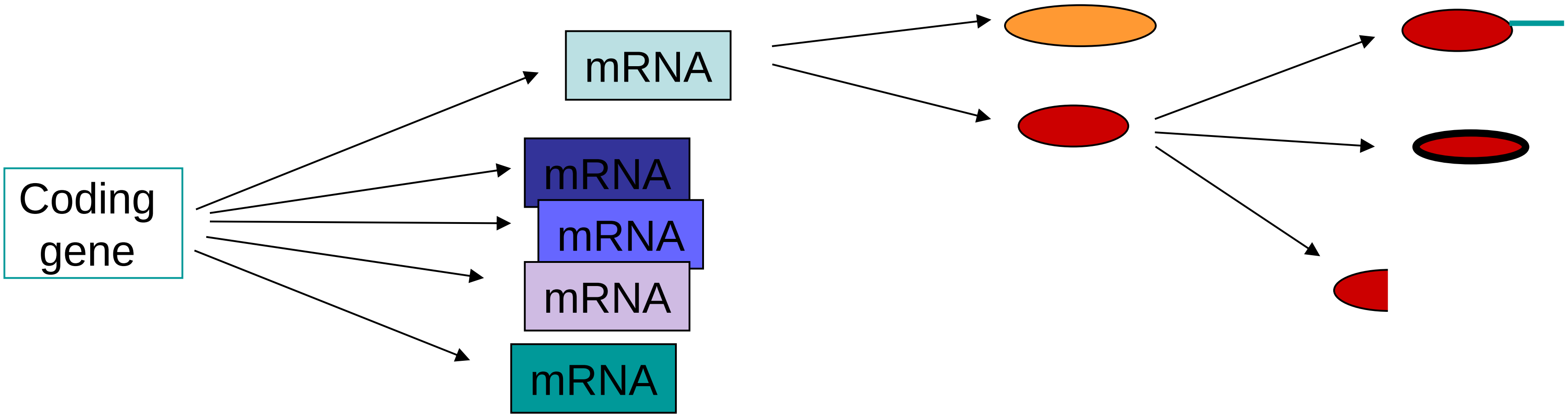
GENERAL CONSIDERATIONS

1) **Number of the proteins**
2) Concentrations of the individual proteins, the dynamic range

Human Proteome – how many proteins are there?

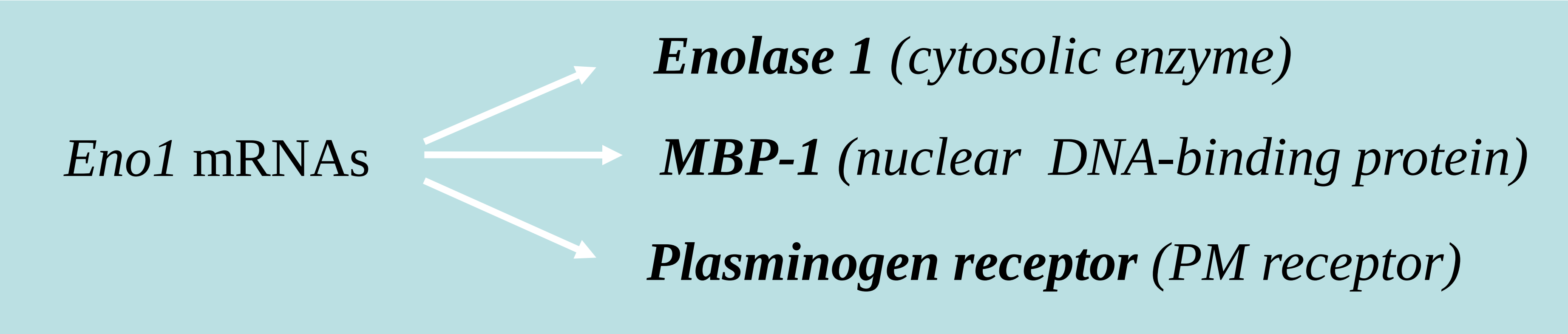
Human Proteome – how many proteins are there?

~20 000 protein-coding genes → **mRNA x 4-8 variants** → **Proteins** **Proteoforms**



WHAT DEFINES A PROTEIN?

The coding gene?
Structure?
Function?



Eno1 mRNAs

Enolase 1 *(cytosolic enzyme)*

MBP-1 *(nuclear DNA-binding protein)*

Plasminogen receptor *(PM receptor)*

One protein or three different proteoforms?

The human proteome – how many proteins do we have?

Gene-centric view

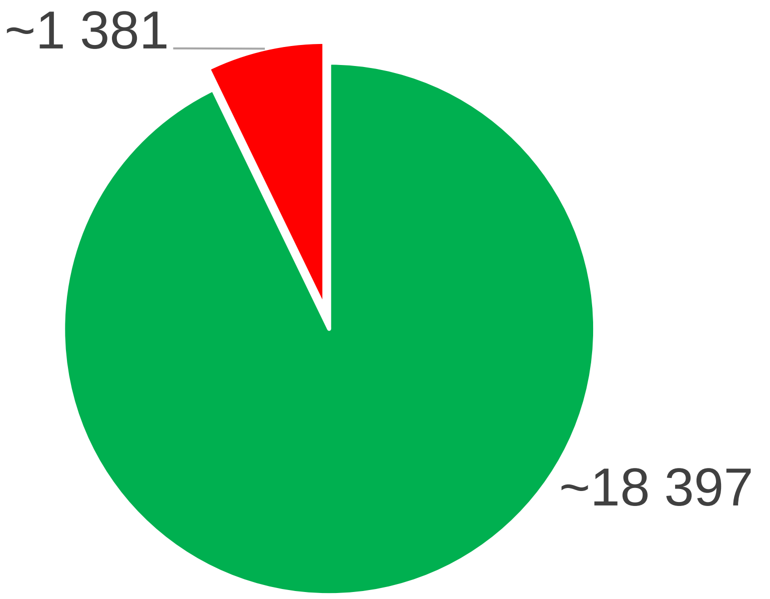
~20 000 protein coding genes → **~ 20 000 proteins**

Proteo-centric view

~20 000 protein coding genes → **> 100 000 proteoforms**

HUMAN PROTEOME PROJECT (STATUS 2023)

19 788 protein-coding genes

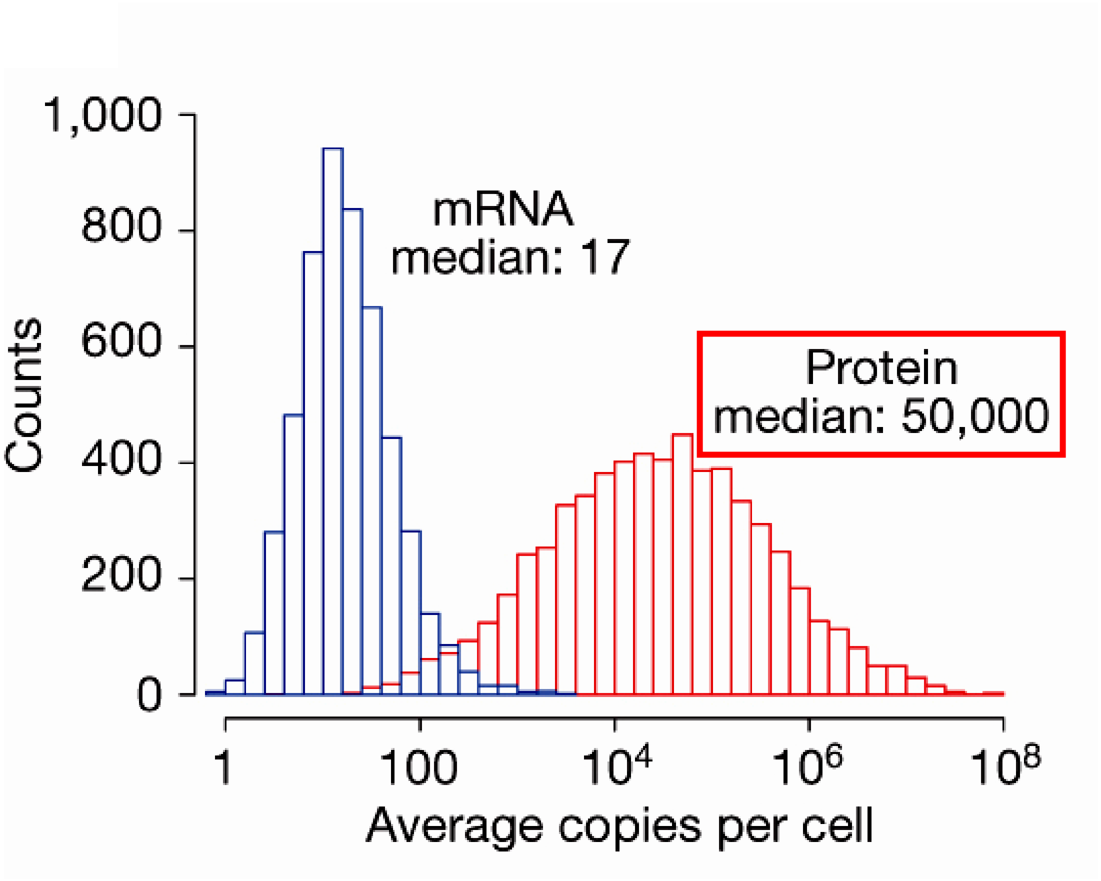


- 🟩 Evidence on protein level (93%)
- 🟥 „The missing proteins“ (7%)

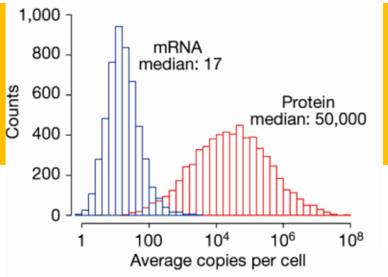
GENERAL CONSIDERATIONS

1) Number of proteins
2) **Concentrations of the individual proteins, the dynamic range**

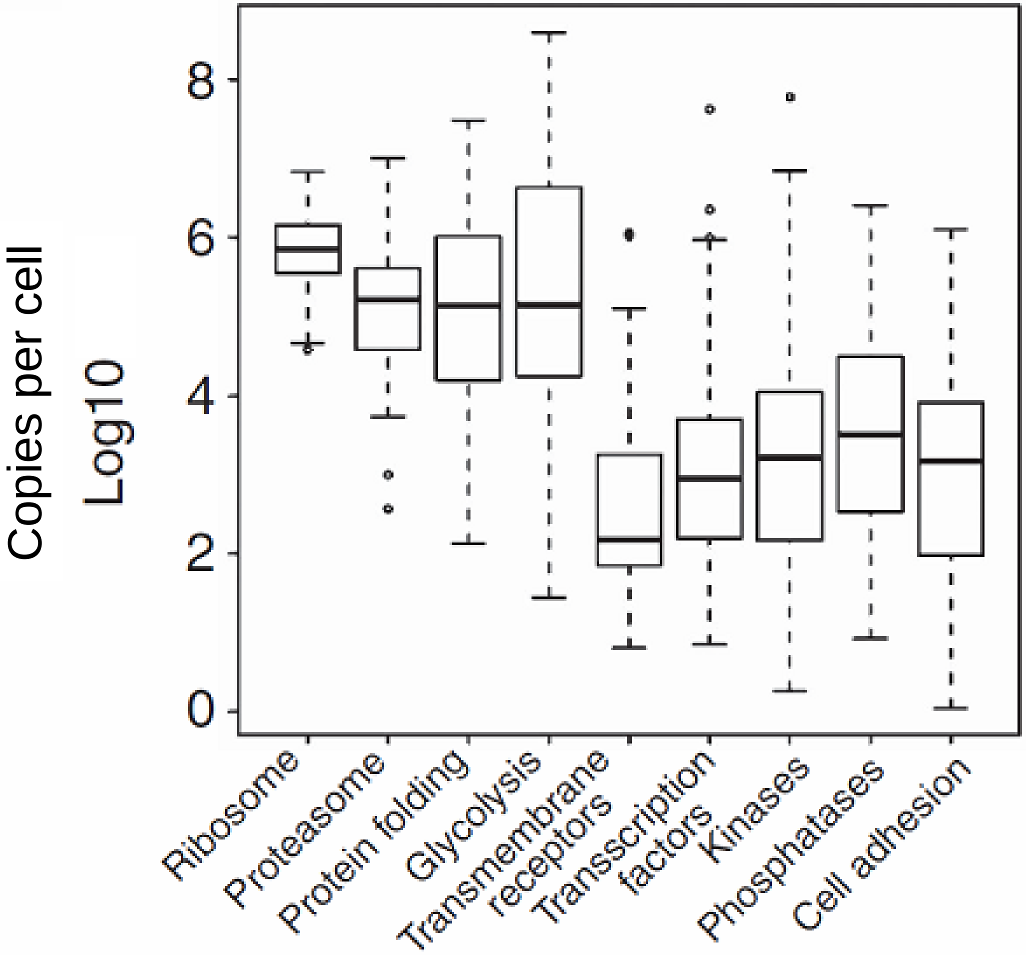
Concentrations of individual cellular proteins (copies/cell) (10-10^8)



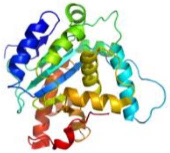
Schwanhäusser B, et al. Nature. 2011;473(7347):337-42



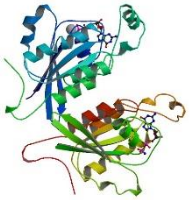
Abundances of cellular proteins
(according the protein function)



Beck et al., Mol Syst Biol. 2011 Nov 8;7:549.



A typical mammalian cell actively transcribes 50-60% of its genes



> 10 000 proteins
Range of concentrations: up to 7- 8 orders of magnitude

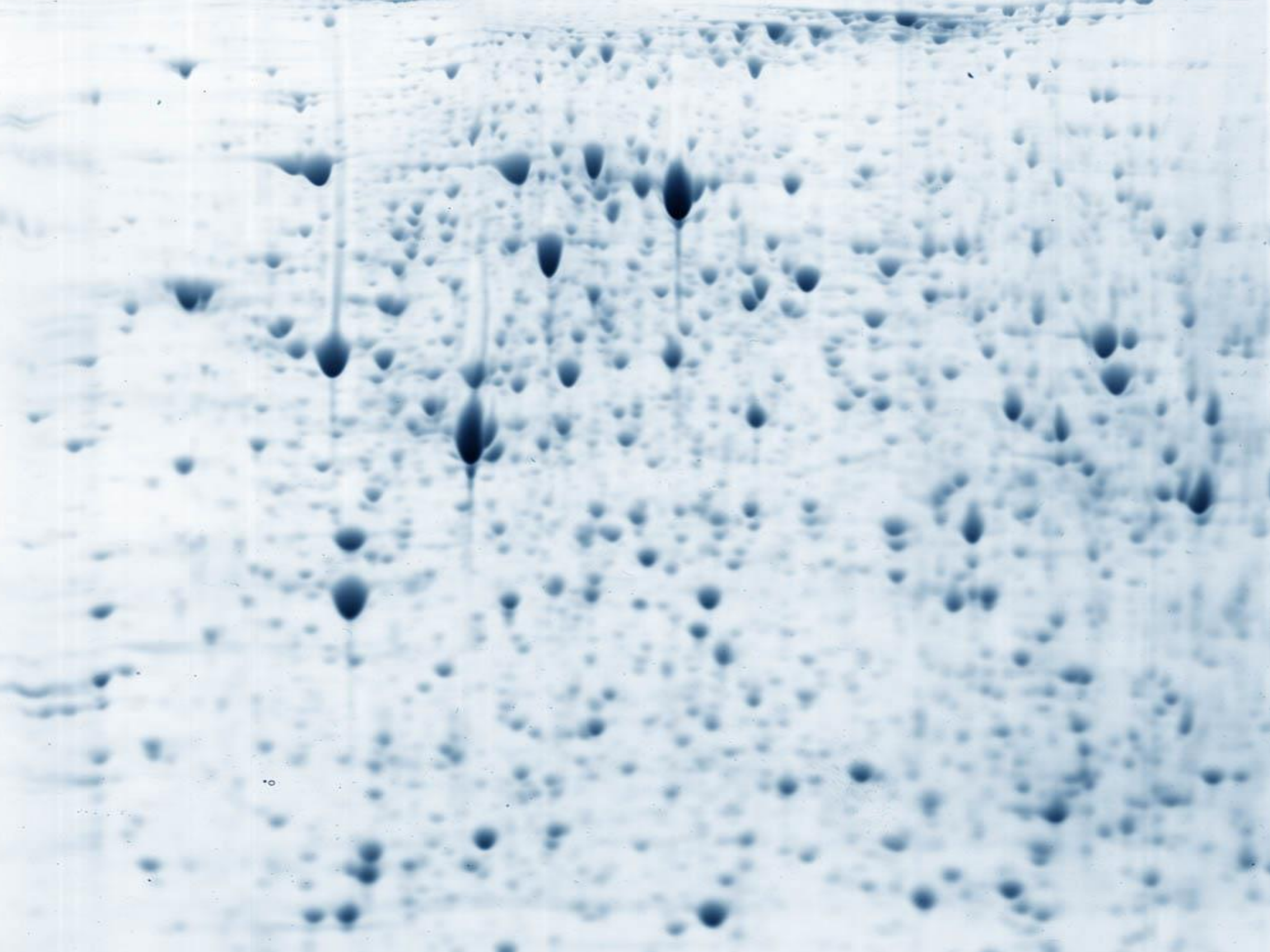
Proteomics uses **liquid chromatography and tandem mass spectrometry** (LC-MS/MS) to identify and quantify as many as possible proteins in the sample

Introduction to proteomics

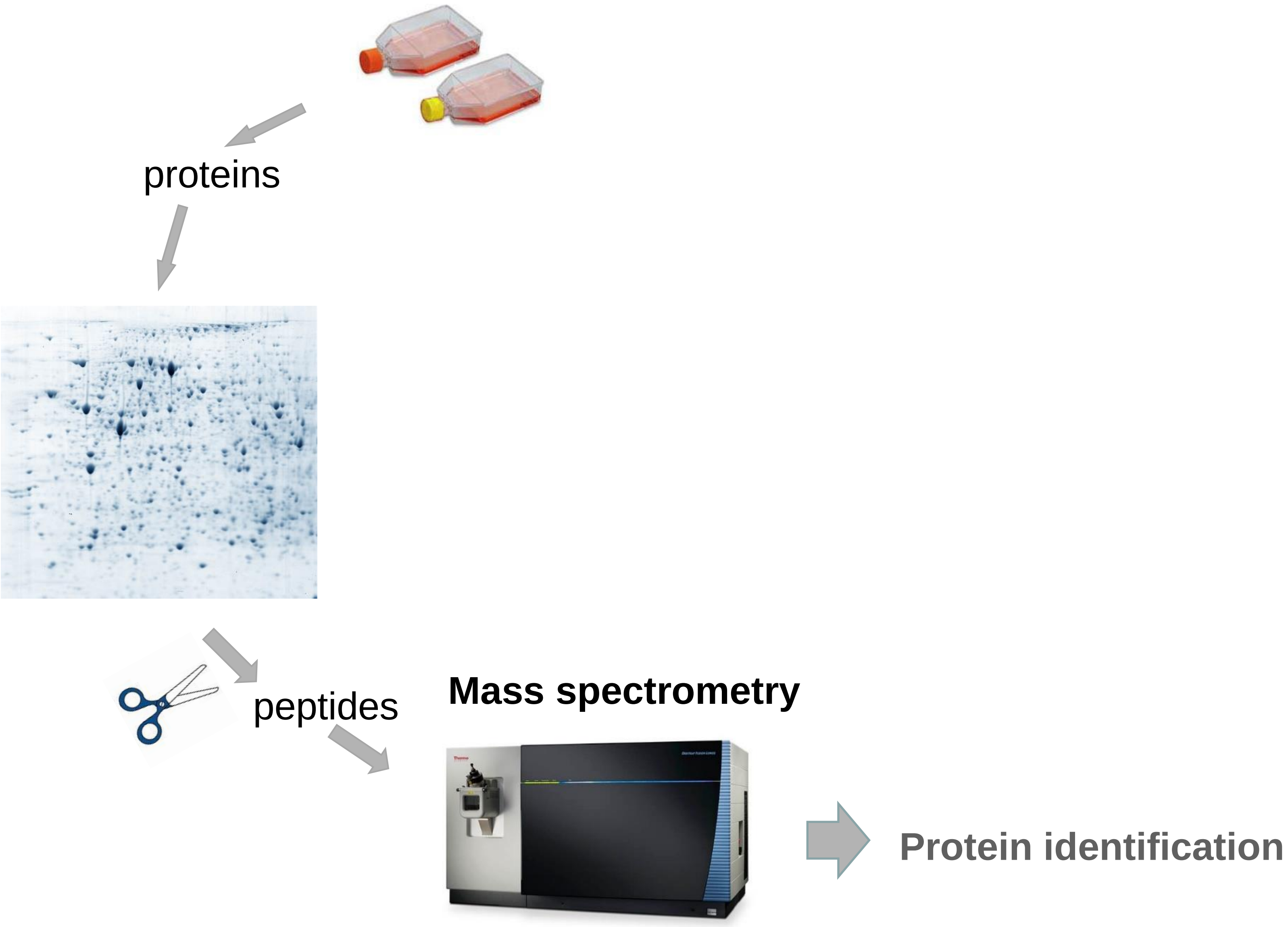
- Human proteome
- Proteomic workflow
- Mass spectrometry and protein identification

An example of a proteomic study

- A quest for circultaing biomarkers of advanced heart failure



HISTORY OF PROTEOMICS
2-DE
(1995-2008)



proteins

peptides

Mass spectrometry

Protein identification

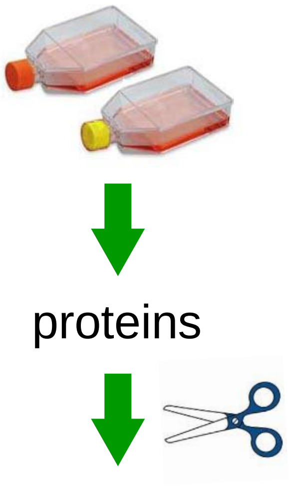
Current Proteomic workflow

LC-MS/MS

proteins

Current Proteomic workflow

LC-MS/MS



proteins

trypsin…

peptides

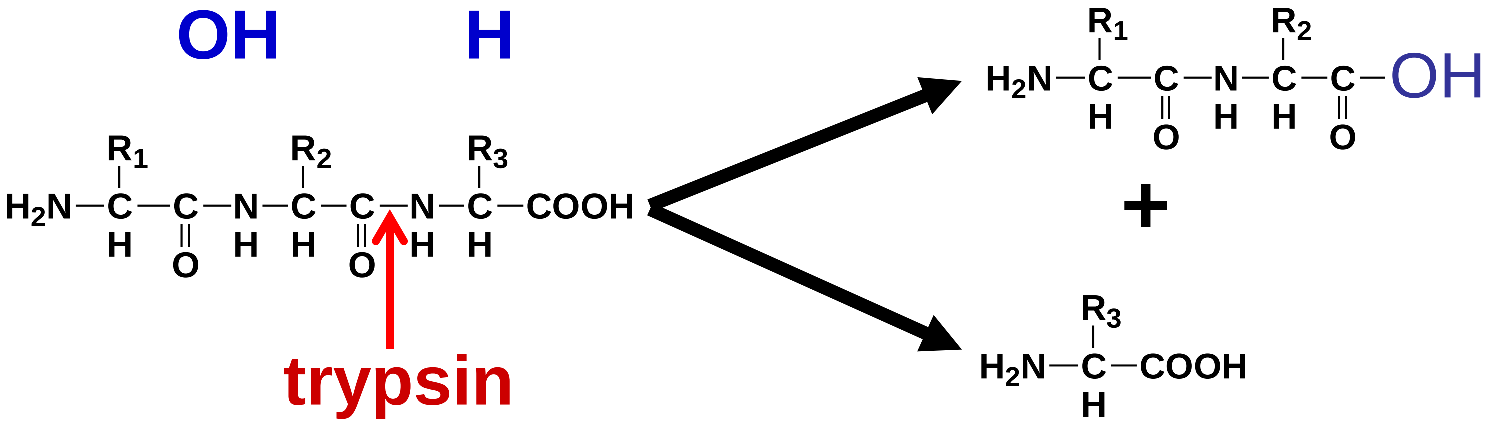
TRYPSIN - a specific endoprotease from pankreas, cleaves after Arg (R) and Lys (K) (unless followed by Pro)

SSQIRQNYSTDVEAAVNSLVNLYLQASYTYLSLGFYFDRDDVALEGVSHFFRELAEEKREG

YERLLKMQNQRGGRALFQD IKKPAEDEWGKTPDAMKAAMALEKKLNQALLDLHAL

GSARTDPHLCDFLETHFLDEEVKL IKKMGDHLTNLHRLGGPEAG LGEYLFERLT LKHD

Digestion with trypsin



Current proteomic workflow
LC-MS/MS

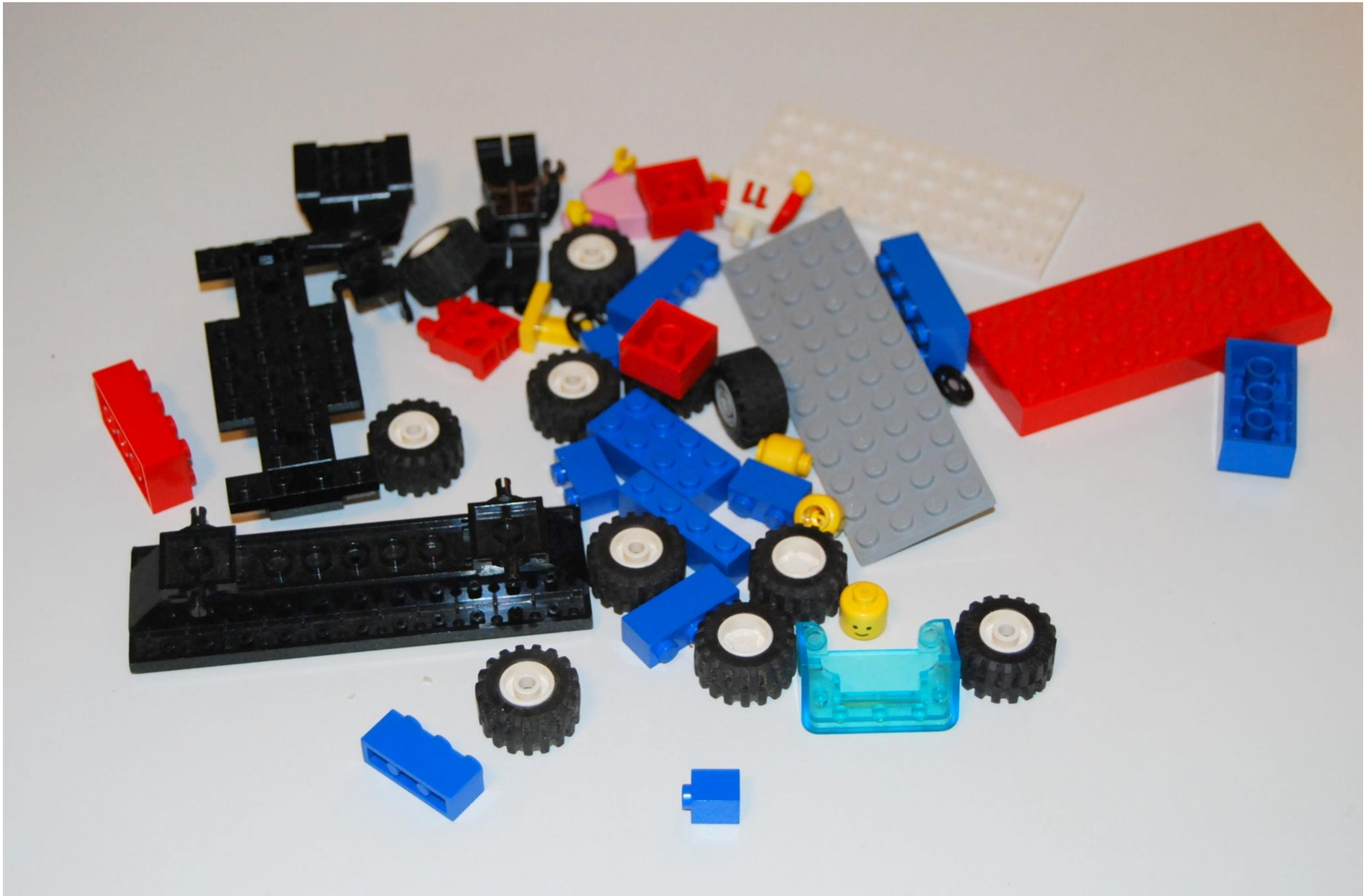
10 000 proteins

TRYPSIN

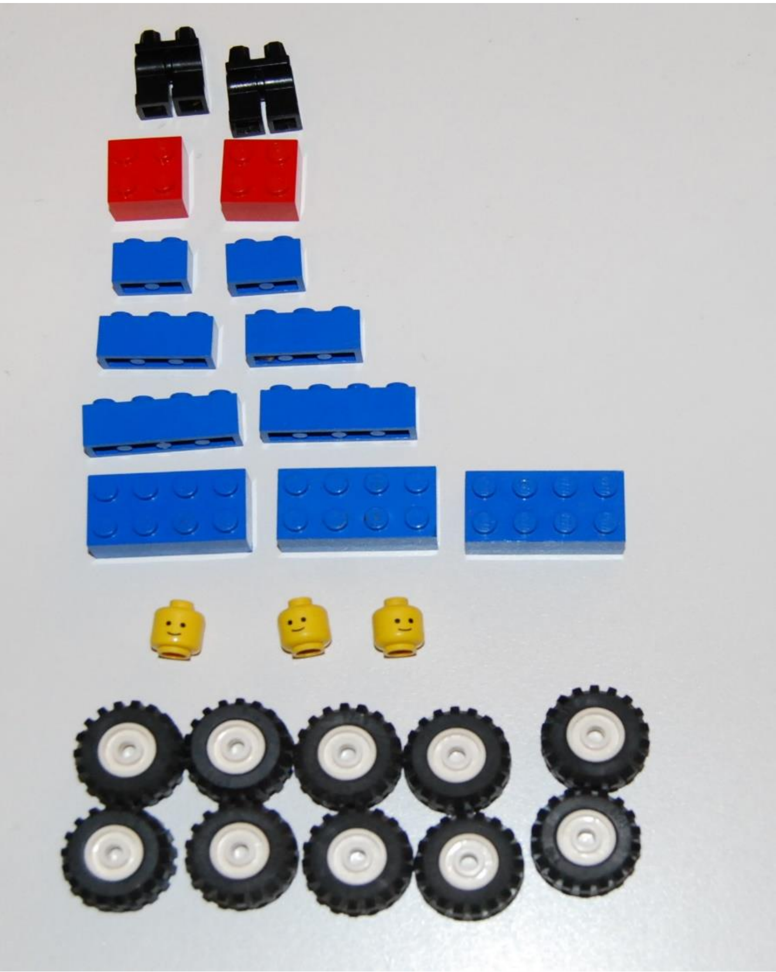
~ 300 000 peptides

- Peptides are more compatible with MS analysis
- Peptides are effectively separated by LC
- To identify a protein we only need a few peptides.

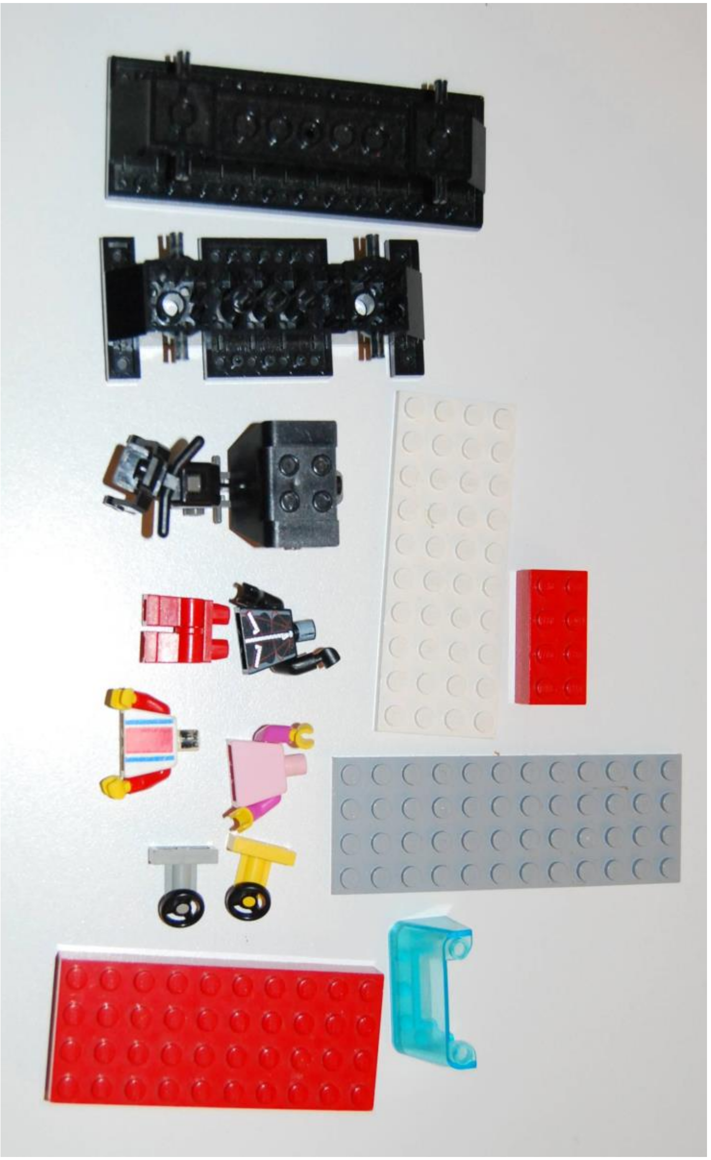
Peptides versus Uniqe peptides



Present in more proteins



Unique



Keratin, type II cytoskeletal 4 (KRT4, P19013)

MIARQQCVRGGPRGFSCGSAIVGGGKRGAFSSVSMSGGAGRCSSGGFGSRSLYNL
RGNKSISMSVAGSRQGACFGGAGGFGTGGFGGGFGGSFSGKGGPGFPVCPAGGI
QEVTINQSLLTPLHVEIDPEIQKVRTEEREQIKLLNNKFASFIDKVQFLEQQNKVLETK
WNLLQQQTTTTSSKNLEPLFETYLSVLRKQLDTLGNDKGRLQSELKTMQDSVEDFK
TKYEEEINKRTAAENDFVVLKKDVDAAYLNKVELEAKVDSLNDEINFLKVLYDAELSQ
MQTHVSDTSVVLSMDNNRNLDLDSIIAEVRAQYEEIAQRSKAEAEALYQTKVQQL
QISVDQHGDNLKNTKSEIAELNRMIQRLRAEIENIKKQCQTLQVSVADAEQRGENA
LKDAHSKRVELEAALQQAKEELARMLREYQELMSVKLALDIEIATYRKLLEGEEYRM
SGECQSAVSISVVSGSTSTGGISGGLGSGSGFGLSSGFGSGSGSGFGFGGSVSGSSS
SKIISTTTLNKRR

Unique tryptic peptides (5-25 AA) represent only 38% in KRT4.

Current proteomic workflow

LC-MS/MS

10 000 proteins

TRYPSIN

~ 300 000 peptides

1D or 2D
Liquid chromatography

Reverse Phase (RP) chromatography

Current proteomic workflow
LC-MS/MS

proteins

trypsin…

peptides

Micro or nano-LC
1D or 2D

Current proteomic workflow

LC-MS/MS

proteins

trypsin…

peptides

Micro or nano-LC
1D or 2D

Mass spectrometry

Peptide/protein identification and quantification

Introduction to proteomics

- Human proteome
- Proteomic workflow
- Mass spectrometry and protein identification

Membrane proteins and proteomics

- Special strategies
- Membrane protome of human neuroendocrine tumors

MASS SPECTROMETRY

Highly accurate measurement of (peptide or peptide fragment) mass (MW).

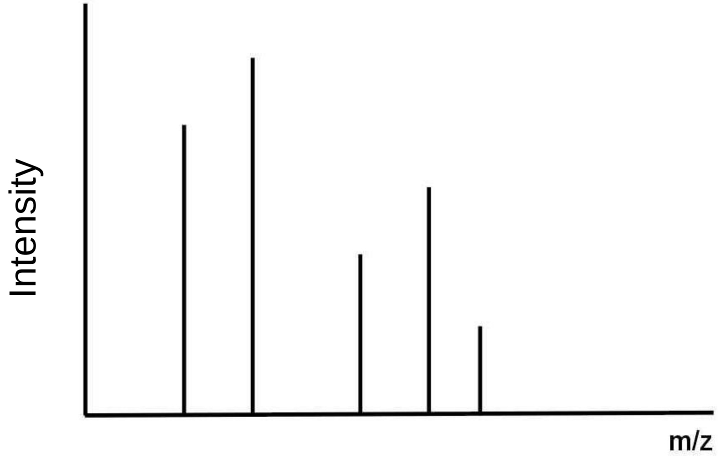
Velocity of a charged particle in electric/magnetic field (in vacuum) is a ***function of its mass*** *and charge (m/z).*

MASS SPECTROMETRY

Highly accurate measurement of (peptide or peptide fragment) mass (MW).

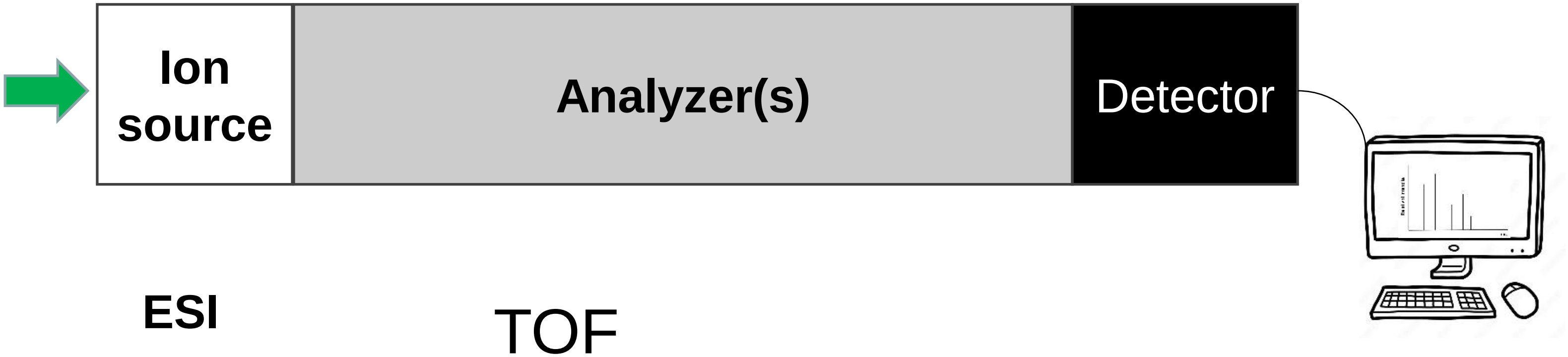
Velocity of a charged particle in electric/magnetic field (in vacuum) is a ***function of its mass*** *and charge (m/z).*

Molecules are ionized and then separated based on their masses (MW). Resulting **mass spectrum** shows relative intensity of individual ions with their respective masses (m/z).



MASS SPECTROMETRY

Mass spectrometer



Ion source

Analyzer(s)

Detector

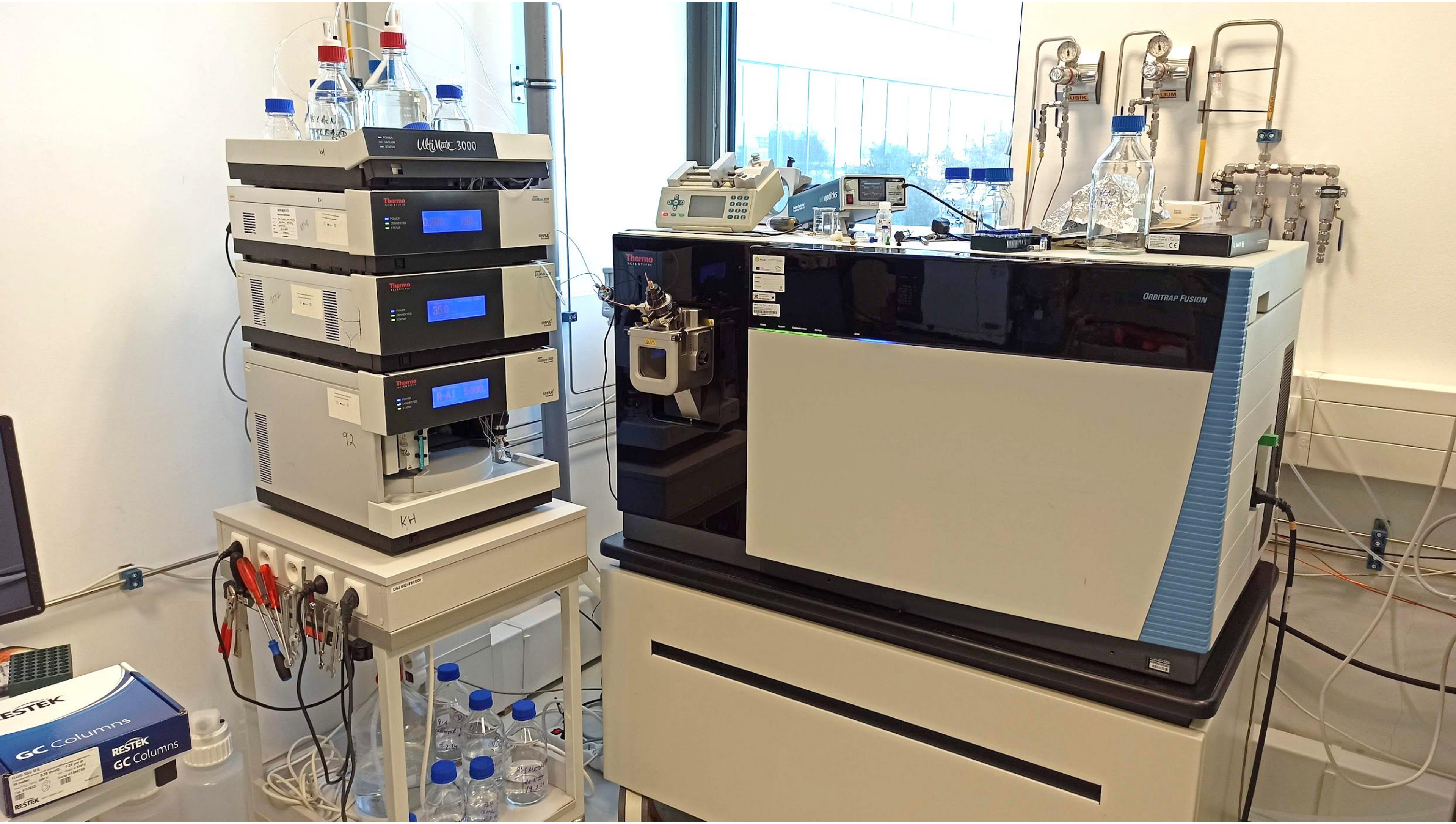
ESI

MALDI

TOF
Quadrupole
Ion trap
Orbital ion trap (Orbitrap)
FTICR
…and their combinations

Analyzers

	mass range	resolution	accuracy	dynamic range
quadrupole	4000	4000	100ppm	1.E+05
Ion Trap	6000	4000	100ppm	1.E+04
TOF	unlimited	50000	5ppm	1.E+04
FT-ICR	10000	>10e7	0.1ppm	1.E+04
Orbitrap	4000	>200000	1ppm	1.E+04

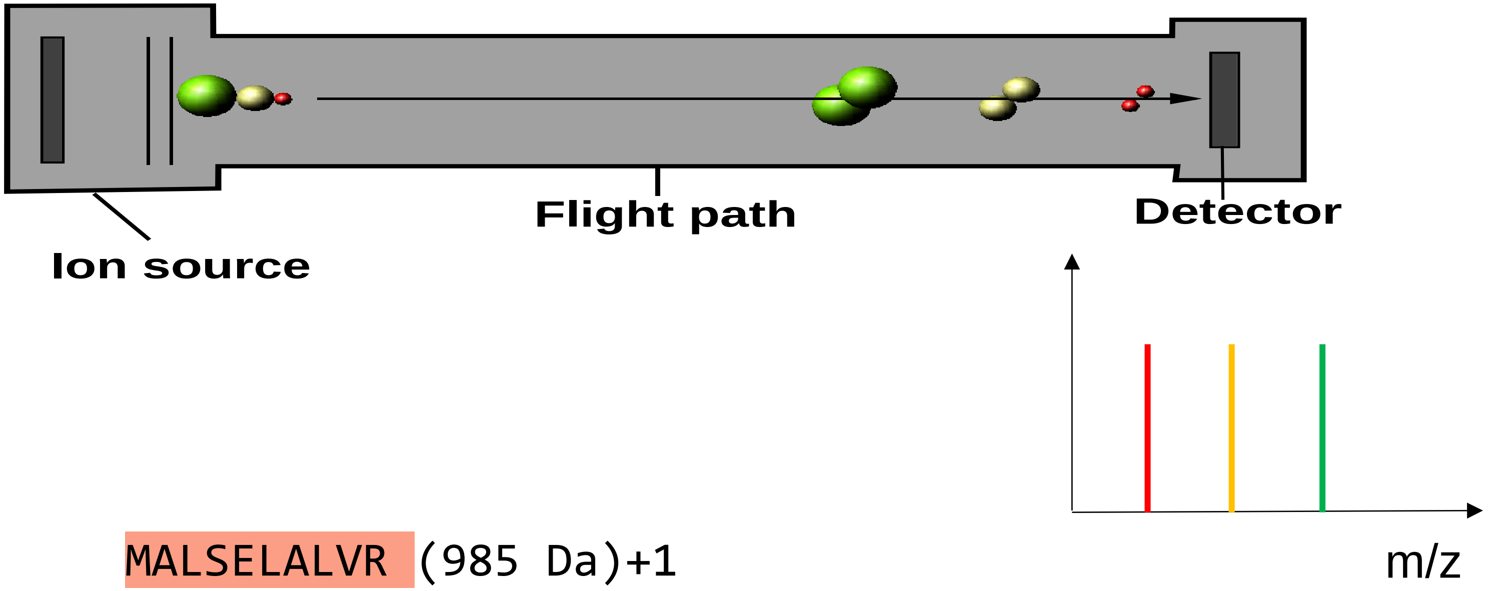

UltiMate 3000
Thermo SCIENTIFIC
ORBITRAP FUSION
RESTEK
GC Columns

Orbitrap MS

Fourier transform ion cyclotron (FTICR) MS

MASS ANALYZER **TIME OF FLIGHT (TOF)**

Linear TOF



MALSELALVR (985 Da)+1

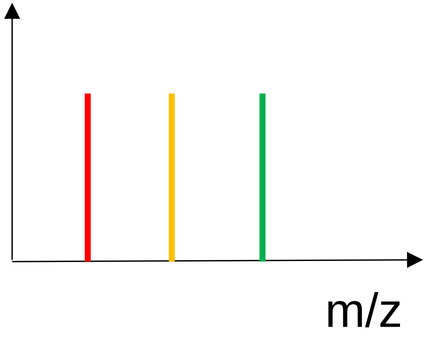
WLQESRRSRKLILFIVFLA (2038 Da)+1

LDNMLLTVVVPIIPSYLYSIKHEKNATEIQTARPVHTSI (3145 Da)+1

MALSELALVR+(985 Da) 985+1/1

WLQESRRSRKLILFIVFLA+ (2038 Da)

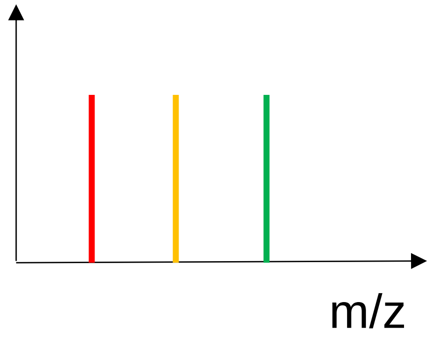
LDNMLLTVVVPIIPSYLYSIKHEKNATEIQTARPVHTSI**+** (3145 Da)



MALSELALVR+(985 Da) 985+1/1

WLQESRRSRKLILFIVFLA+ (2038 Da)

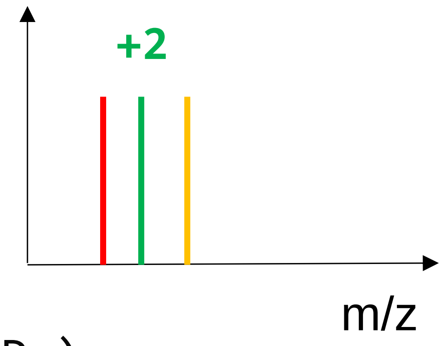
LDNMLLTVVVPIIPSYLYSIKHEKNATEIQTARPVHTSI+ (3145 Da)



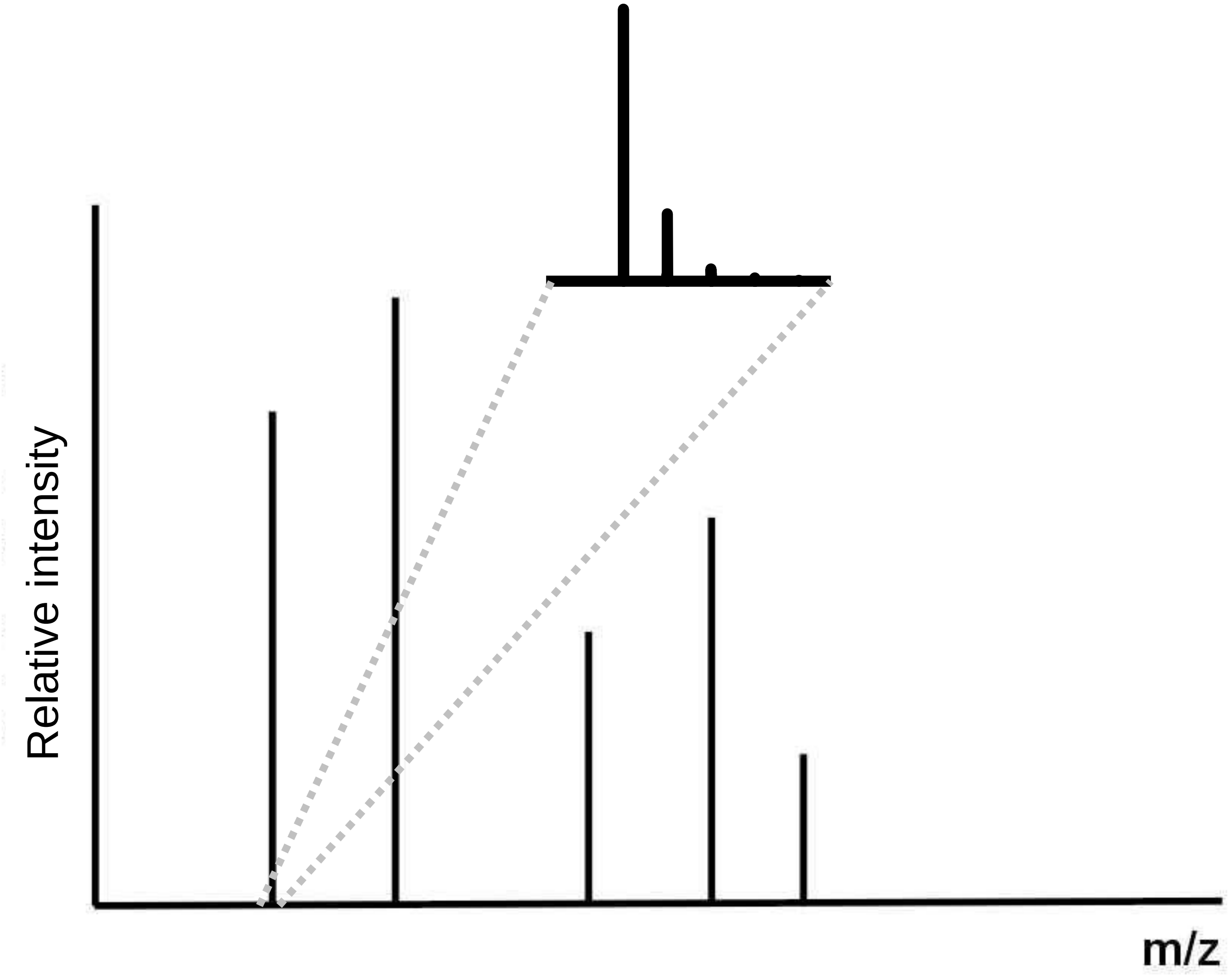
MALSELALVR+(985 Da)

WLQESRRSRKLILFIVFLA+ (2038 Da)

LDNMLLTVVVPIIPSYLYSIKHEKNATEIQTARPVHTSI ++ (3145 Da)
(3145+2)/2=1573



Mass spectrum in detail – isotopic cluster



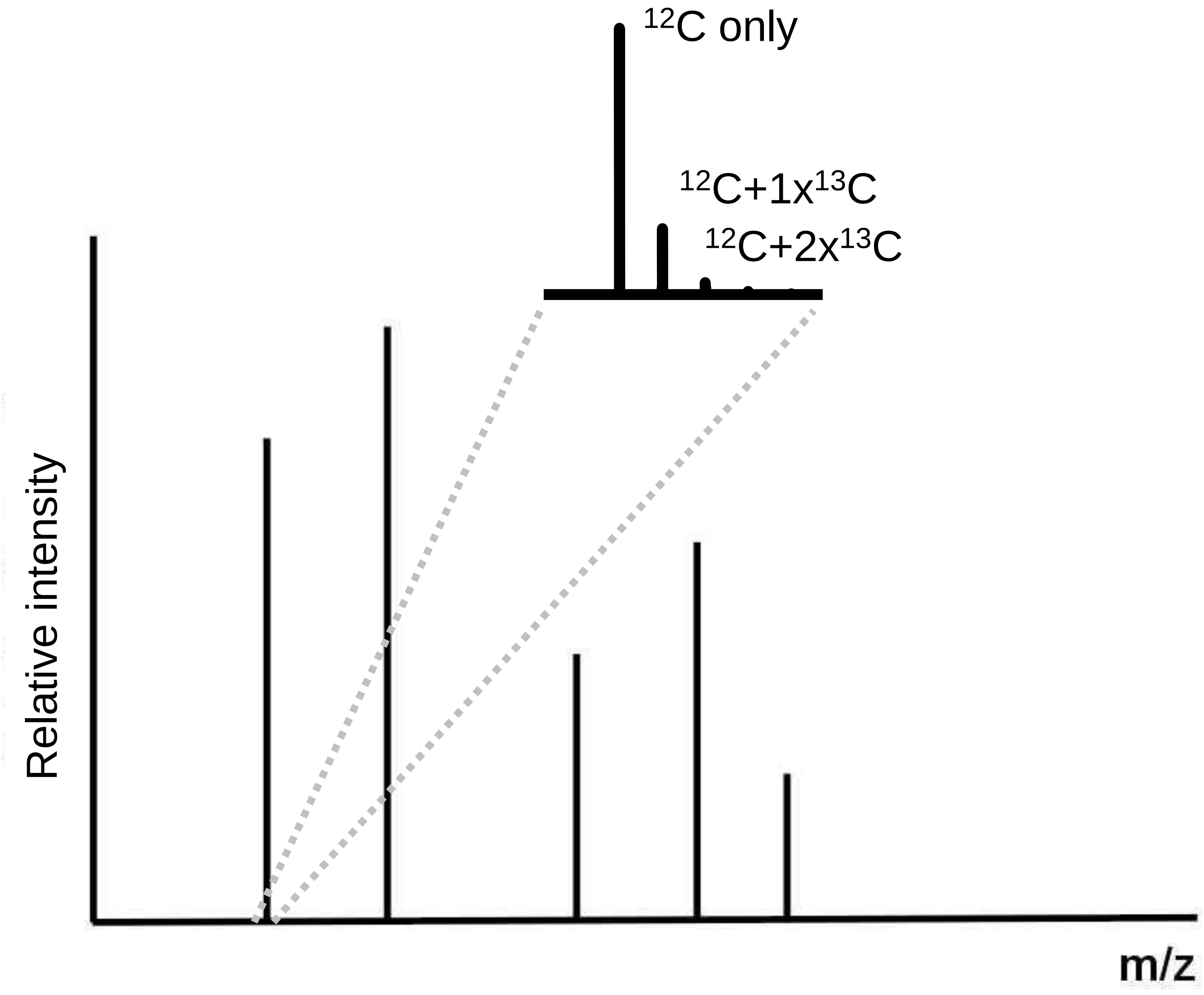
Courtesy of Petr Man

Masses and isotopes

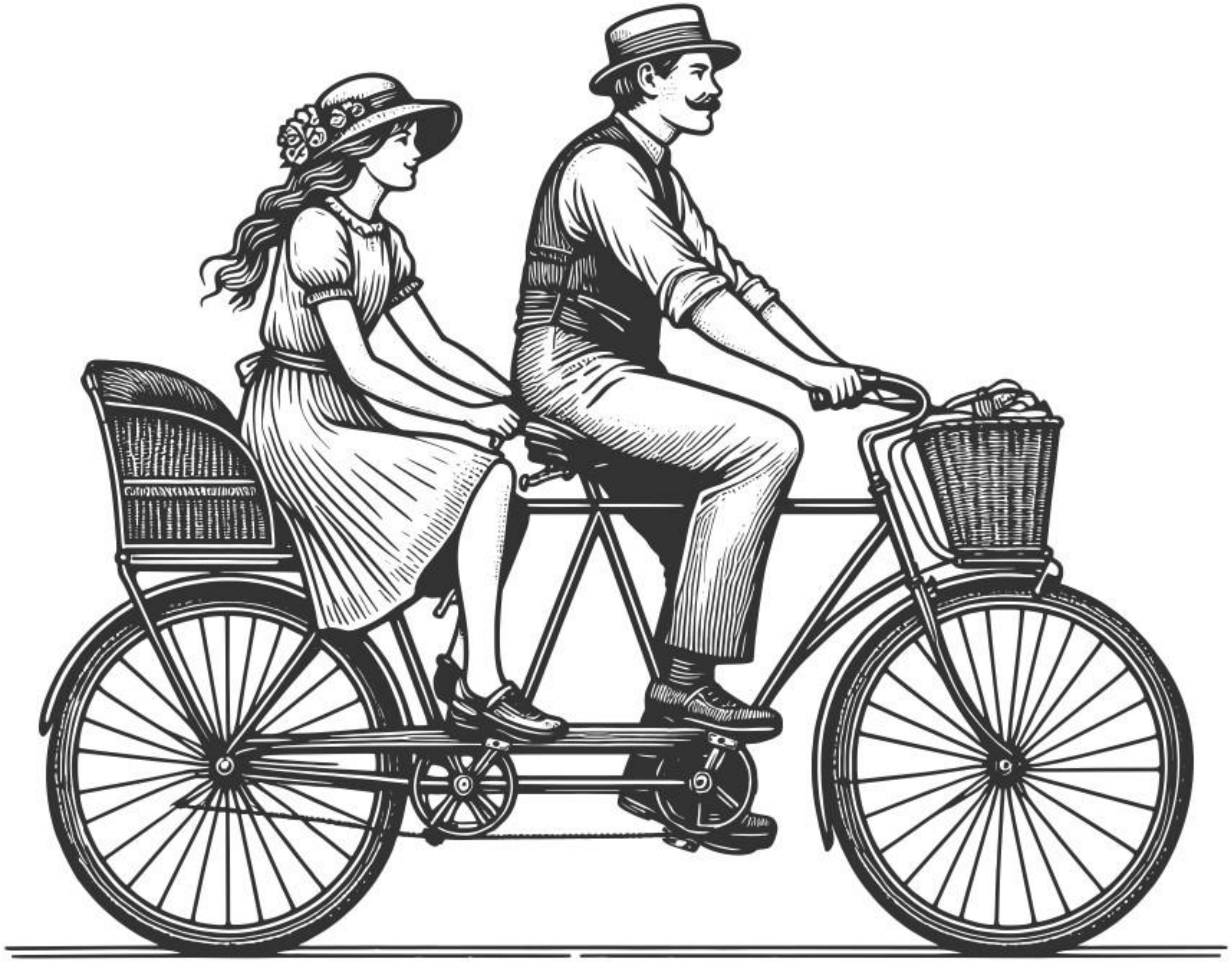
Symbol	M_{nom}	M_{mono}	%
C	12	12.00000	98.9300
	13	13.00336	1.0700
H	1	1.00783	99.9885
	2	2.01411	0.1150
N	14	14.00307	99.6320
	15	15.00011	0.3680
O	16	15.99492	99.7570
	17	16.99913	0.0380
	18	17.99916	0.2050
S	32	31.97207	94.9300
	33	32.97146	0.7600
	34	33.96787	4.2900
	36	35.96708	0.0200
P	31	30.97376	100.0000

Courtesy of Petr Man

Mass spectrum in detail – isotopic cluster



Courtesy of Petr Man



PROTEIN IDENTIFICATION BY **TANDEM MASS SPECTROMETRY** (MS/MS)

PROTEIN IDENTIFICATION BY **TANDEM MASS SPECTROMETRY** (MS/MS)

1. Masses of **peptides** are measured by MS

PROTEIN IDENTIFICATION BY **TANDEM MASS SPECTROMETRY** (MS/MS)

1. Masses of **peptides** are measured by MS

2. Individual peptide is isolated and **fragmented** inside the spectrometer

PROTEIN IDENTIFICATION BY **TANDEM MASS SPECTROMETRY** (MS/MS)

1. Masses of **peptides** are measured by MS
2. Individual peptide is isolated and **fragmented** inside the spectrometer
3. Masses of the resulting **peptide fragments** are measured by MS

PROTEIN IDENTIFICATION BY **TANDEM MASS SPECTROMETRY** (MS/MS)

1. Masses of **peptides** are measured by MS

2. Individual peptide is isolated and **fragmented** inside the spectrometer

3. Masses of the resulting **peptide fragments** are measured by MS

The determined masses are **compared to theoretical masses of all peptides or peptide fragments** coded by all genes present in a gene/protein database.

Proteomic workflow
LC-MS/MS

proteins

trypsin…

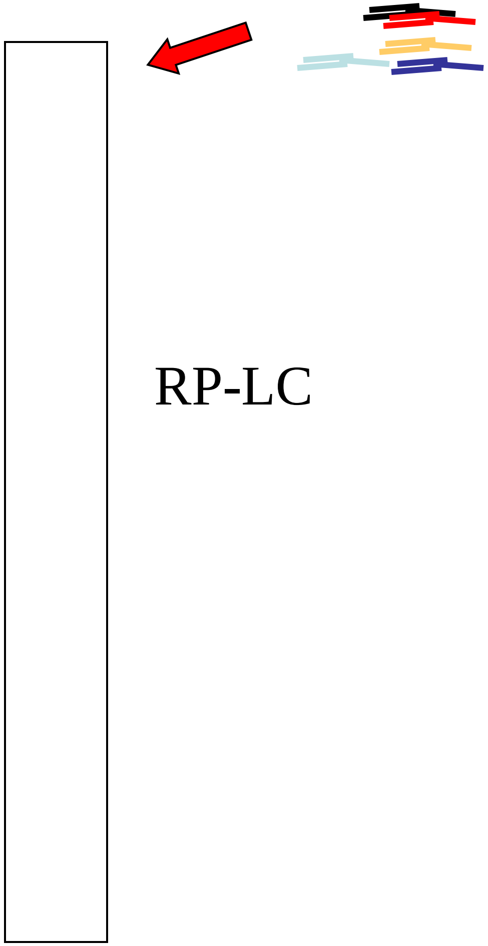
peptides

ON-LINE

Mass spectrometry

Peptide/protein identification and quantification

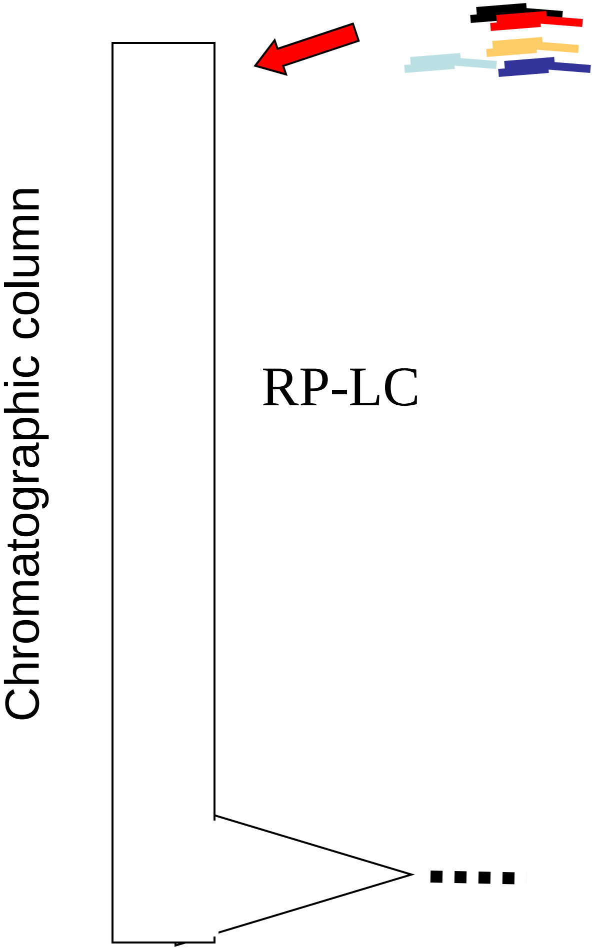
LC-MS/MS



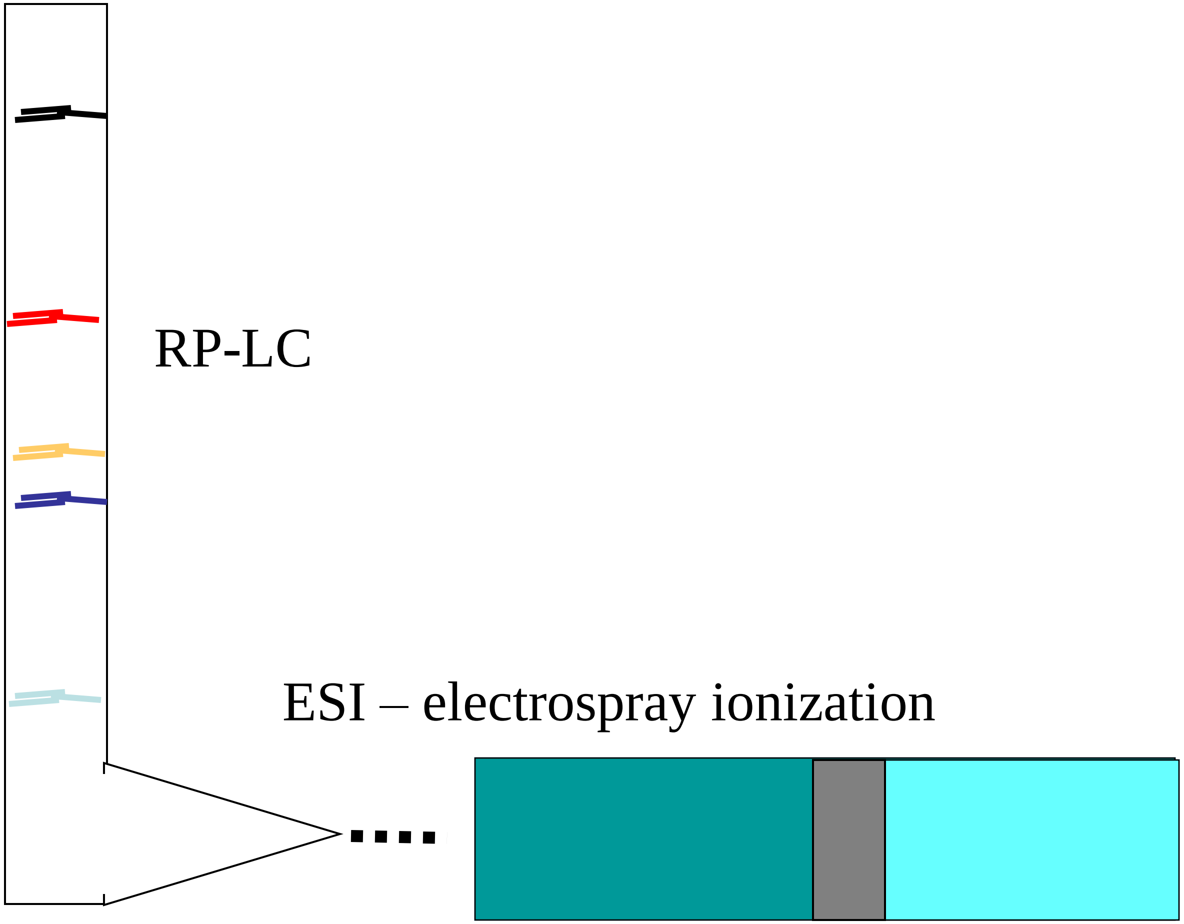
Chromatographic column

RP-LC

LC-MS/MS



LC-MS/MS



RP-LC

ESI – electrospray ionization

MS1

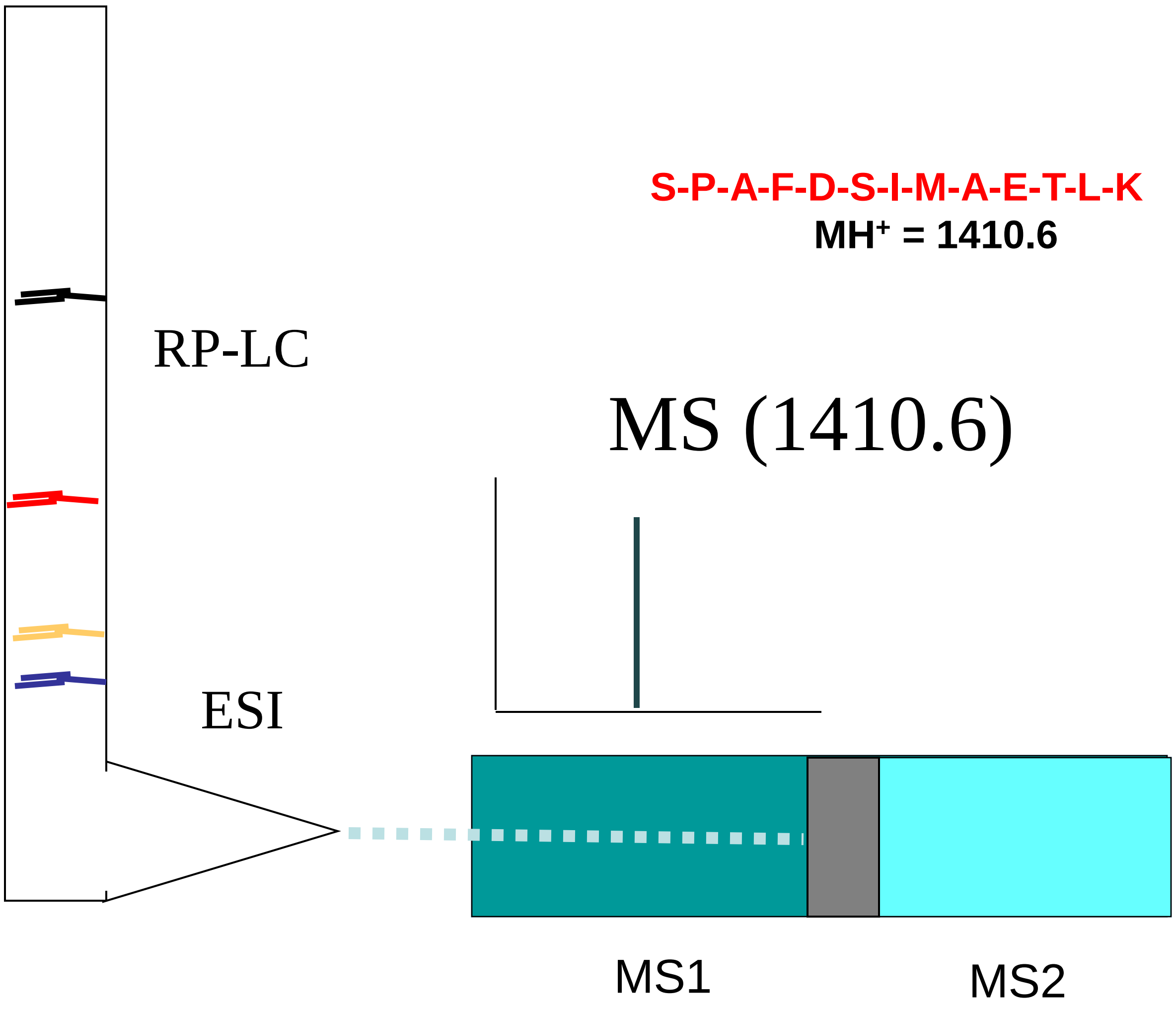
MS2

MS/MS (tandem) mass spectrometer

LC-MS/MS

1. Determine the accurate **mass of the peptide**

LC-MS/MS



S-P-A-F-D-S-I-M-A-E-T-L-K

$MH^+ = 1410.6$

RP-LC

MS (1410.6)

ESI

MS1

MS2

MS/MS Mass spectrometer

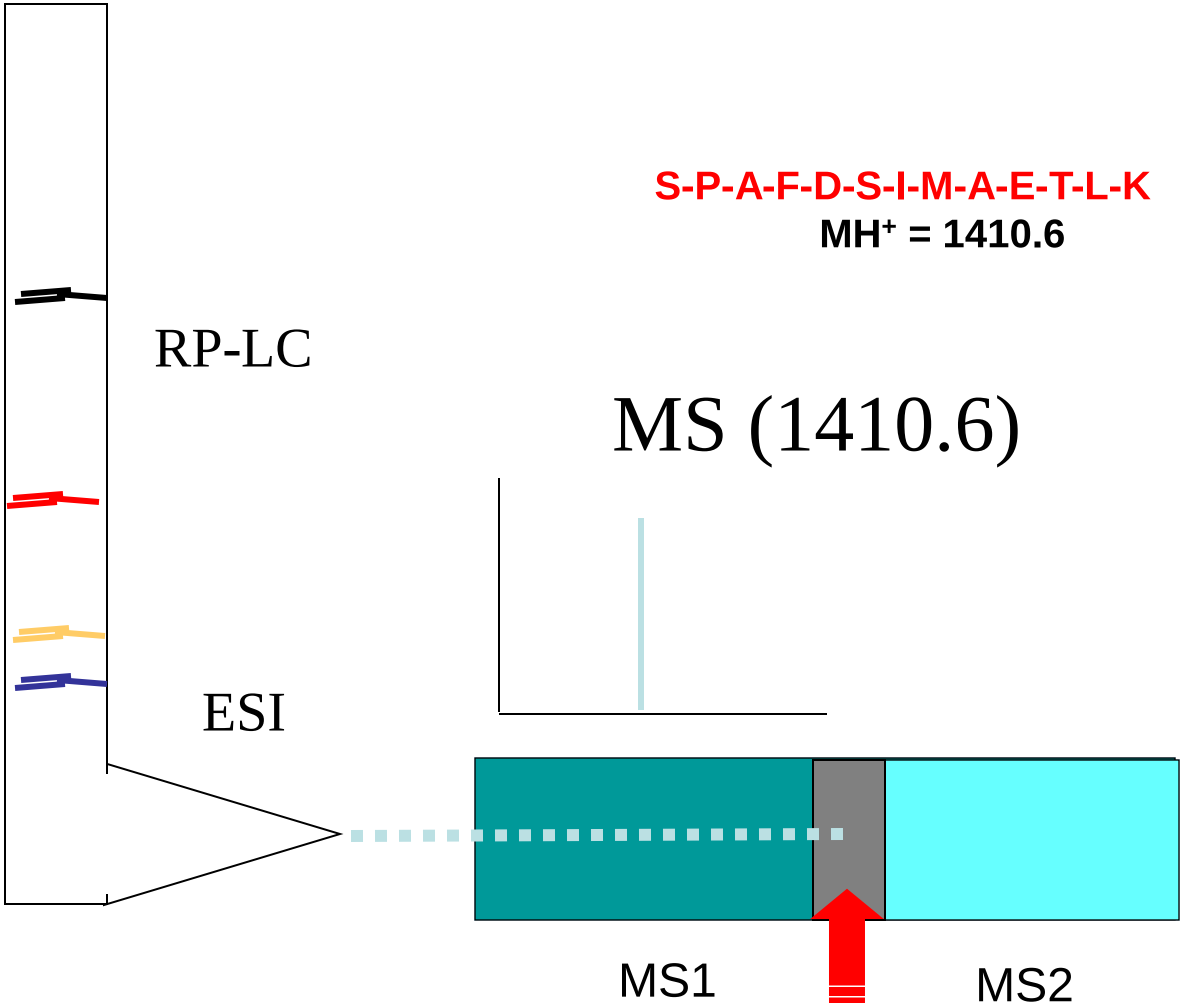
LC-MS/MS

1. Determine the accurate **mass of the peptide**
2. Fragment the peptide inside the MS

LC-MS/MS

S-P-A-F-D-S-I-M-A-E-T-L-K

$MH^+ = 1410.6$



RP-LC

MS (1410.6)

ESI

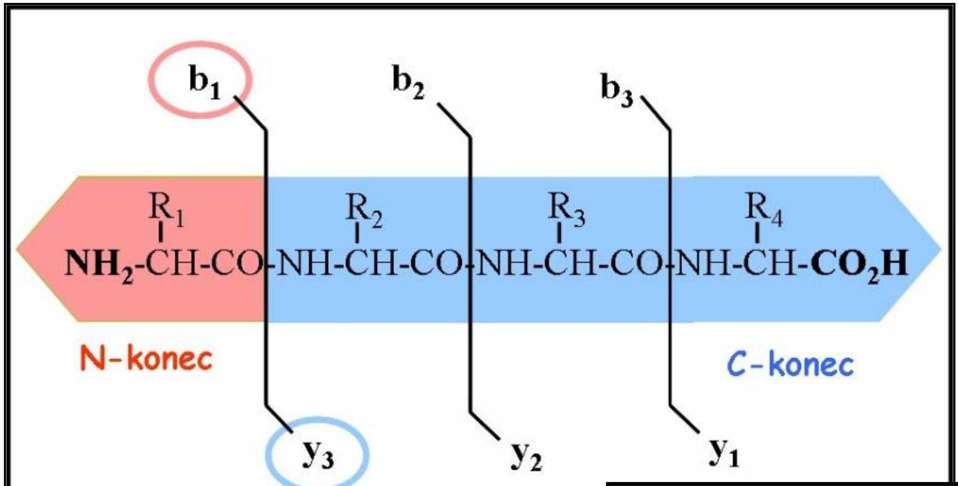
MS1

MS2

Collision cell

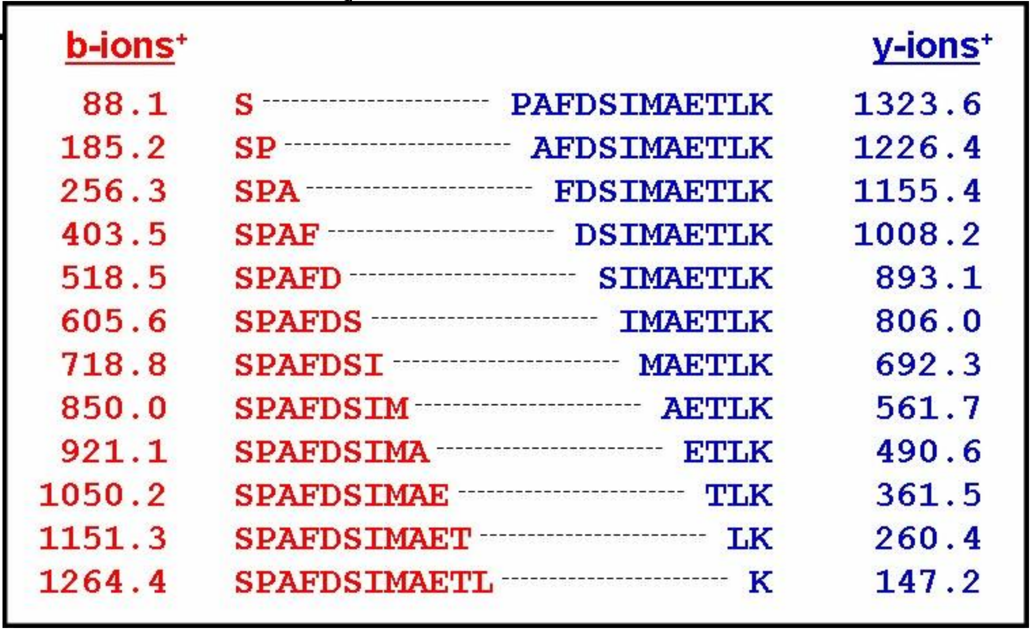
(inert gas, laser, transfer of electrons…)

Peptide fragmentation (microsequencing, MS/MS)



S-P-A-F-D-S-I-M-A-E-T-L-K

MH$^+$ = 1410.6



b-ions$^+$			y-ions$^+$
88.1	S	PAFDSIMAETLK	1323.6
185.2	SP	AFDSIMAETLK	1226.4
256.3	SPA	FDSIMAETLK	1155.4
403.5	SPAF	DSIMAETLK	1008.2
518.5	SPAFD	SIMAETLK	893.1
605.6	SPAFDS	IMAETLK	806.0
718.8	SPAFDSI	MAETLK	692.3
850.0	SPAFDSIM	AETLK	561.7
921.1	SPAFDSIMA	ETLK	490.6
1050.2	SPAFDSIMAE	TLK	361.5
1151.3	SPAFDSIMAET	LK	260.4
1264.4	SPAFDSIMAETL	K	147.2

Courtesy of P. Mann

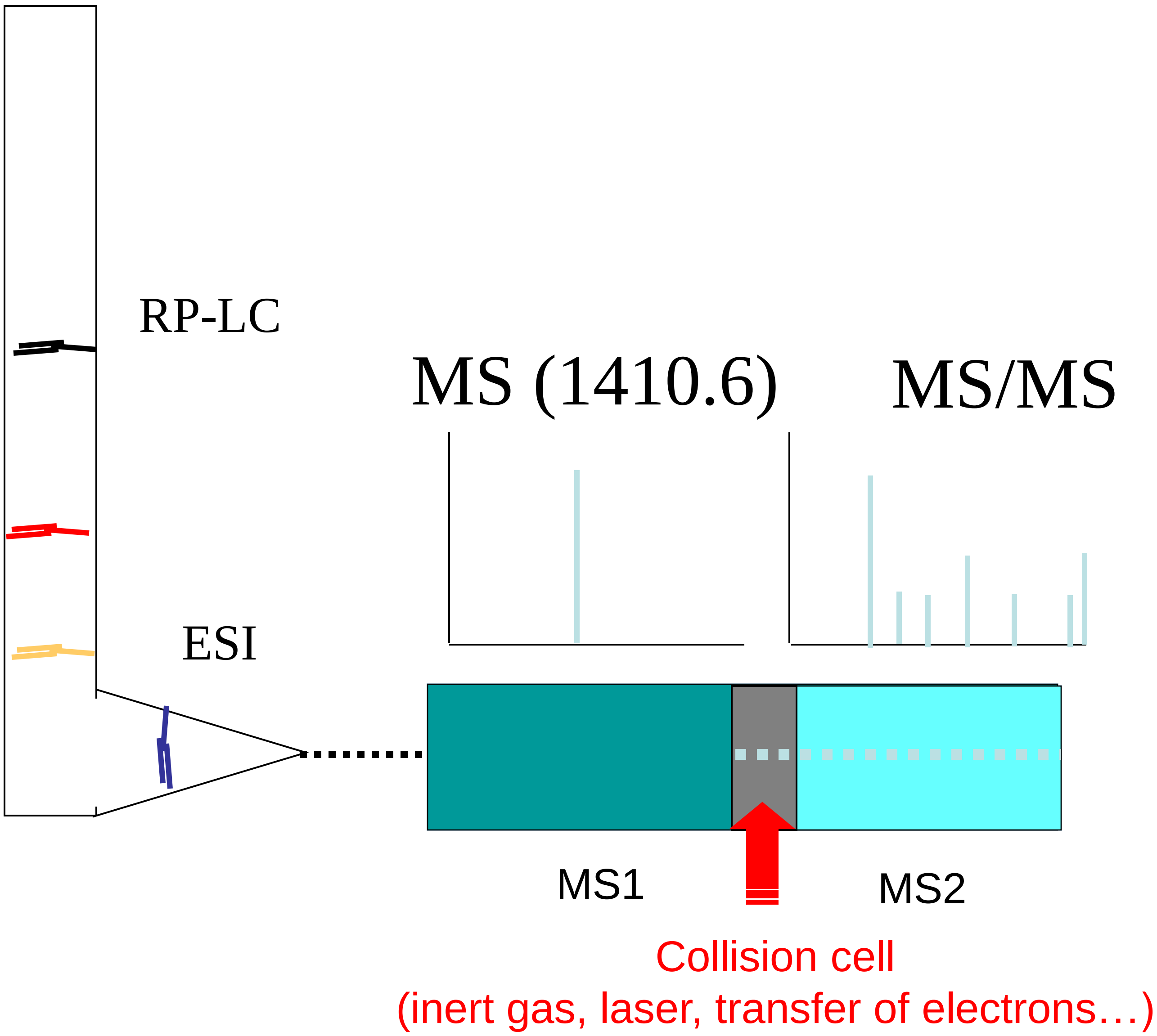
LC-MS/MS

1. Determine the accurate **mass of the peptide**

2. Fragment the peptide inside MS

3. Determine the **masses of the resulting peptide fragments**

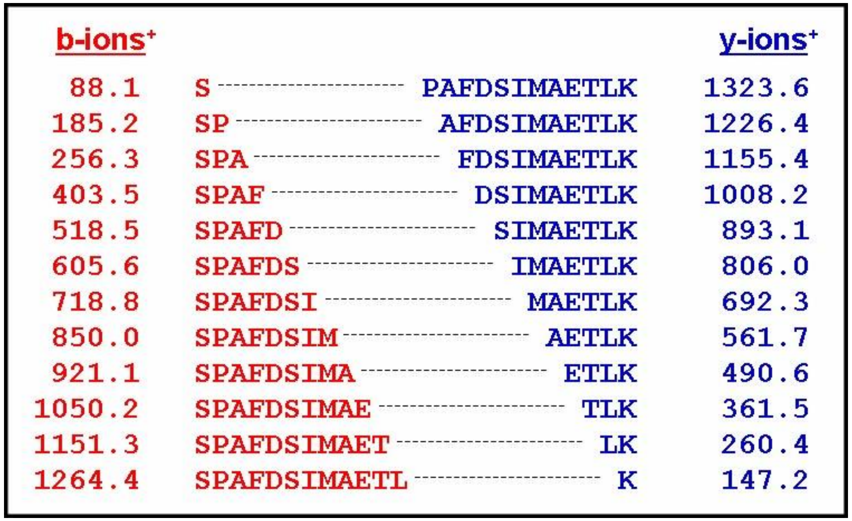
LC-MS/MS


RP-LC
ESI
MS (1410.6)
MS/MS
MS1
MS2
Collision cell
(inert gas, laser, transfer of electrons…)

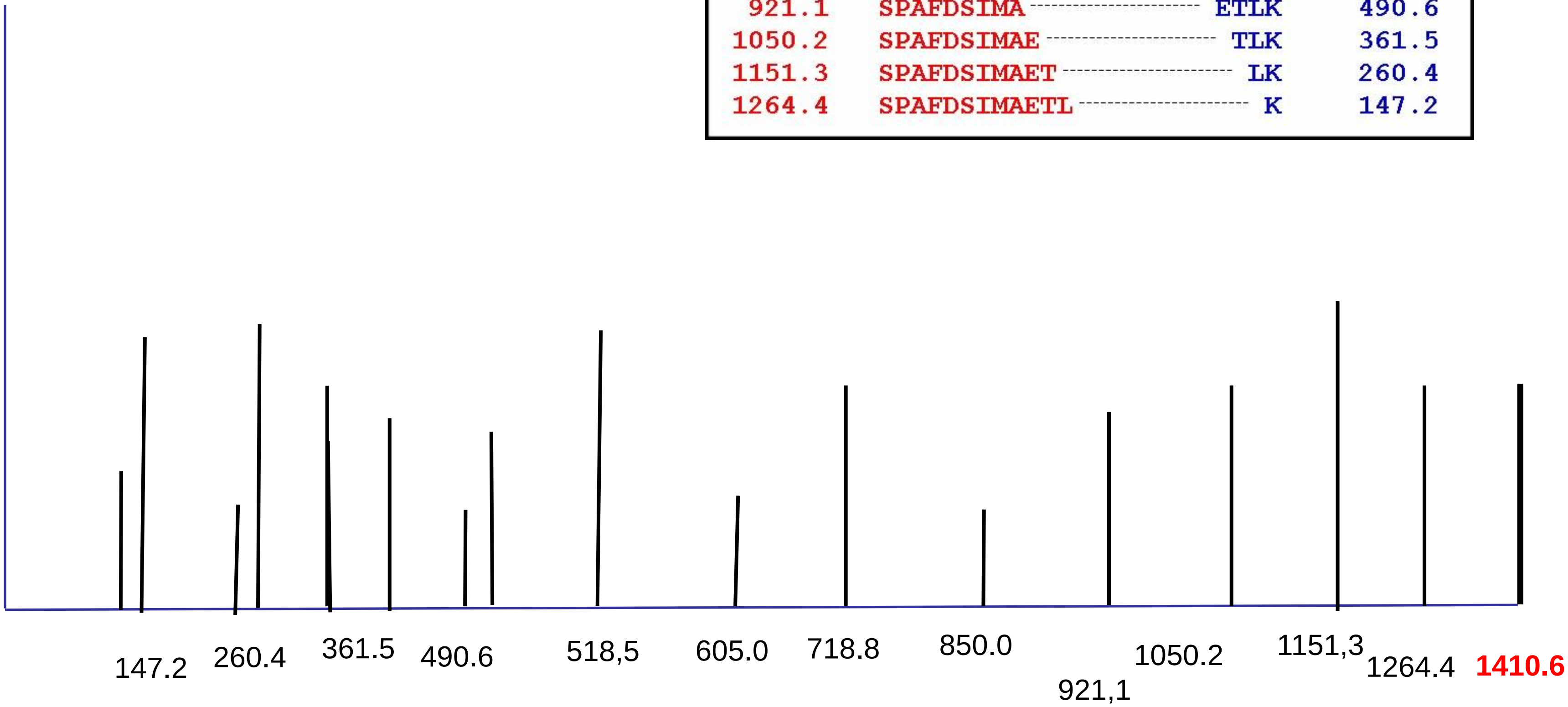
MS2

S-P-A-F-D-S-I-M-A-E-T-L-K

MH⁺ = 1410.6



b-ions⁺			y-ions⁺
88.1	S	PAFDSIMAETLK	1323.6
185.2	SP	AFDSIMAETLK	1226.4
256.3	SPA	FDSIMAETLK	1155.4
403.5	SPAF	DSIMAETLK	1008.2
518.5	SPAFD	SIMAETLK	893.1
605.6	SPAFDS	IMAETLK	806.0
718.8	SPAFDSI	MAETLK	692.3
850.0	SPAFDSIM	AETLK	561.7
921.1	SPAFDSIMA	ETLK	490.6
1050.2	SPAFDSIMAE	TLK	361.5
1151.3	SPAFDSIMAET	LK	260.4
1264.4	SPAFDSIMAETL	K	147.2

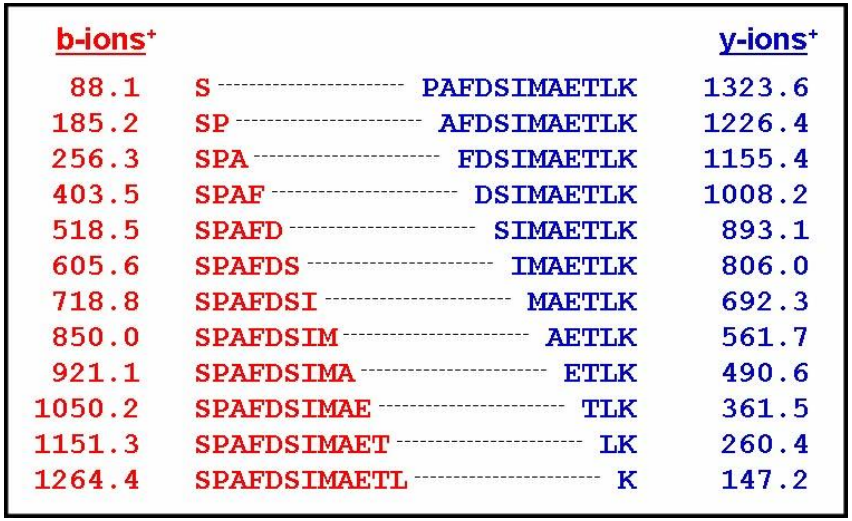


Amino acid fragments

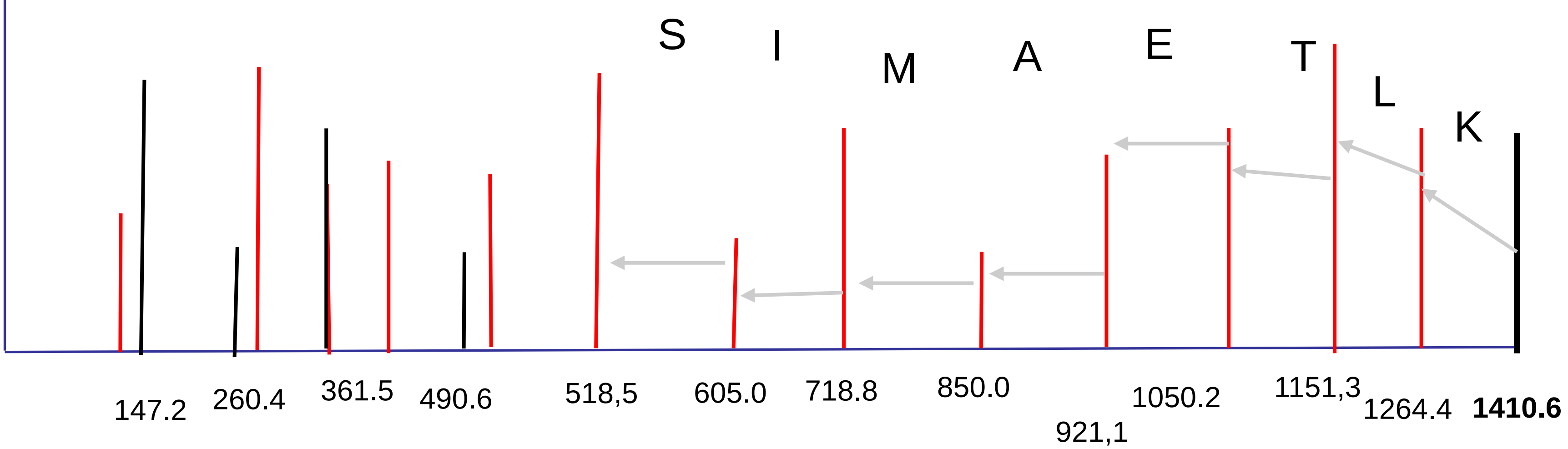
AA		Mass
G	Gly	57.02
A	Ala	71.08
S	Ser	87.03
P	Pro	97.05
V	Val	99.07
T	Thr	101.05
C	Cys	103.01
L	Leu	113.08
I	Ile	113.08
N	Asn	114.04
D	Asp	115.03
Q	Gln	128.06
K	Lys	128.09
E	Glu	129.04
M	Met	131.04
H	His	137.06
F	Phe	147.07
R	Arg	156.10
Y	Tyr	163.06
W	Trp	186.08

S-P-A-F-D-S-I-M-A-E-T-L-K

MH$^+$ = 1410.6

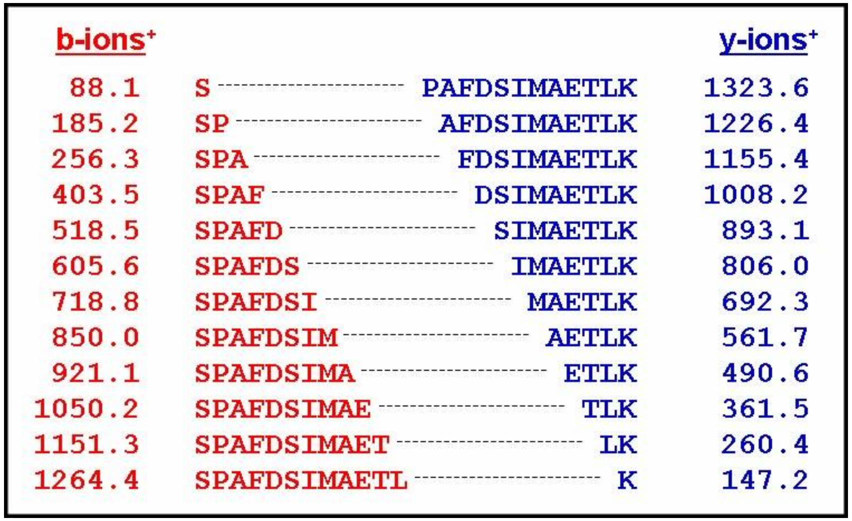


b-ions$^+$			y-ions$^+$
88.1	S	PAFDSIMAETLK	1323.6
185.2	SP	AFDSIMAETLK	1226.4
256.3	SPA	FDSIMAETLK	1155.4
403.5	SPAF	DSIMAETLK	1008.2
518.5	SPAFD	SIMAETLK	893.1
605.6	SPAFDS	IMAETLK	806.0
718.8	SPAFDSI	MAETLK	692.3
850.0	SPAFDSIM	AETLK	561.7
921.1	SPAFDSIMA	ETLK	490.6
1050.2	SPAFDSIMAE	TLK	361.5
1151.3	SPAFDSIMAET	LK	260.4
1264.4	SPAFDSIMAETL	K	147.2

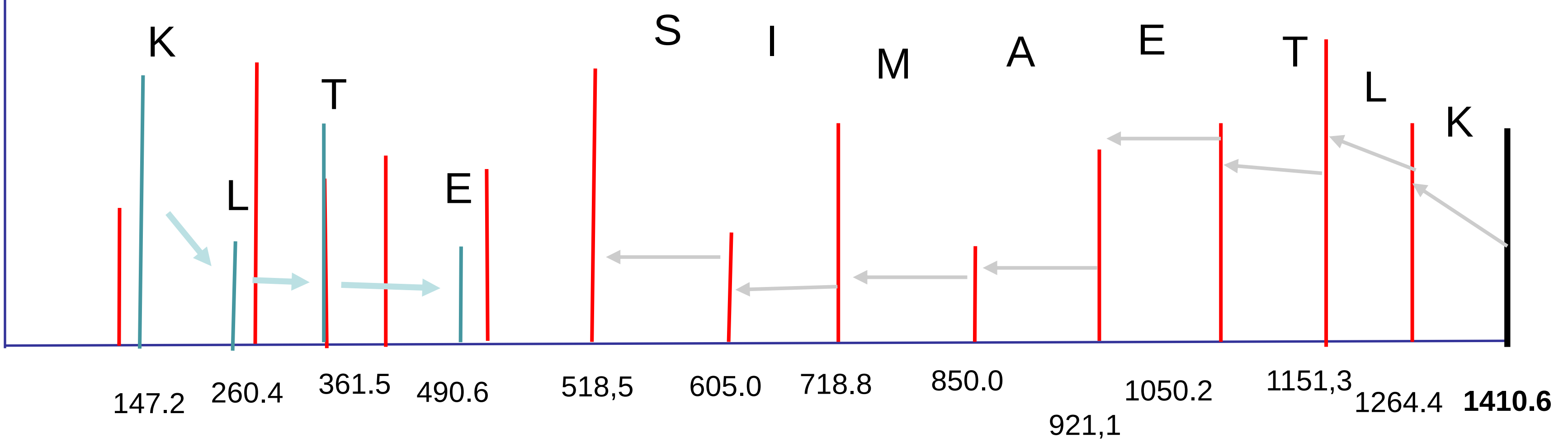


S-P-A-F-D-S-I-M-A-E-T-L-K

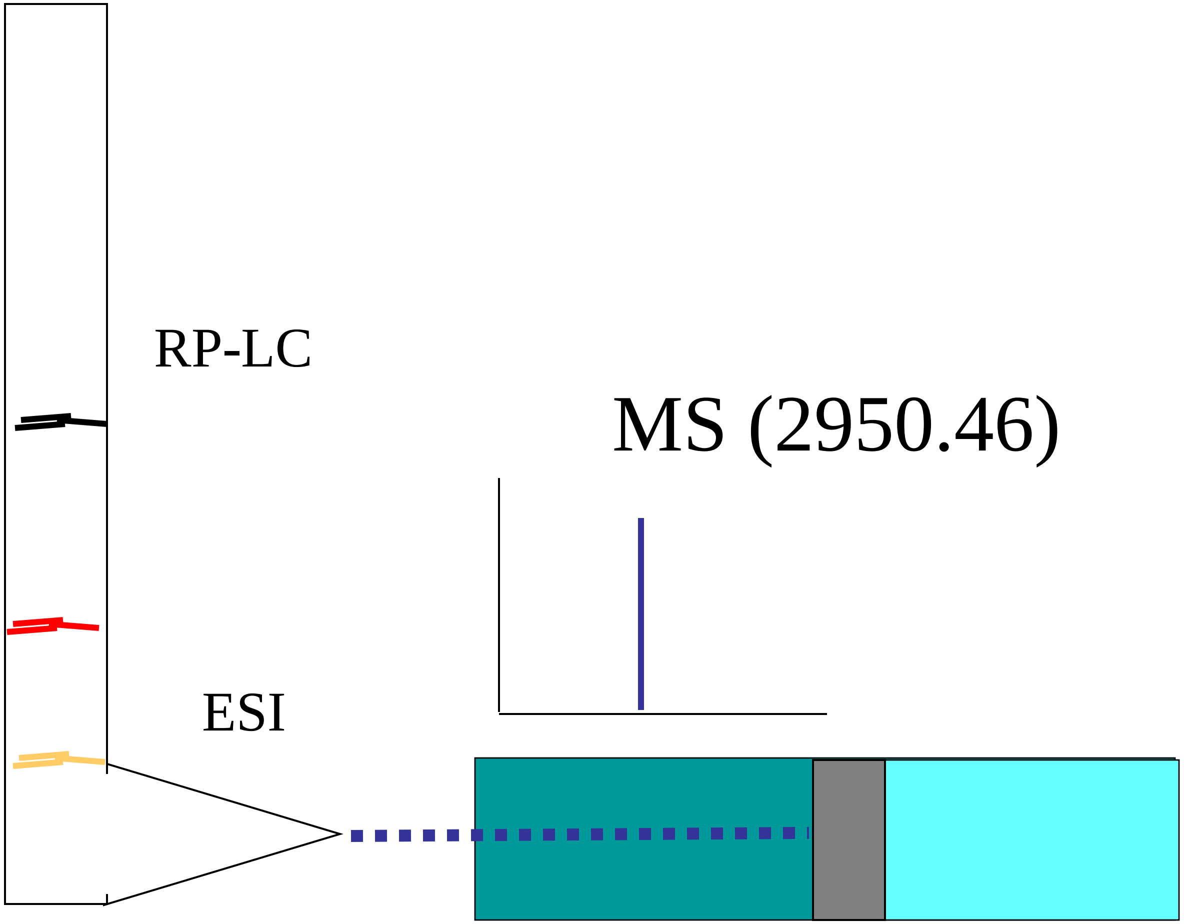
MH$^+$ = 1410.6



b-ions$^+$			y-ions$^+$
88.1	S	PAFDSIMAETLK	1323.6
185.2	SP	AFDSIMAETLK	1226.4
256.3	SPA	FDSIMAETLK	1155.4
403.5	SPAF	DSIMAETLK	1008.2
518.5	SPAFD	SIMAETLK	893.1
605.6	SPAFDS	IMAETLK	806.0
718.8	SPAFDSI	MAETLK	692.3
850.0	SPAFDSIM	AETLK	561.7
921.1	SPAFDSIMA	ETLK	490.6
1050.2	SPAFDSIMAE	TLK	361.5
1151.3	SPAFDSIMAET	LK	260.4
1264.4	SPAFDSIMAETL	K	147.2



LC-MS/MS



RP-LC

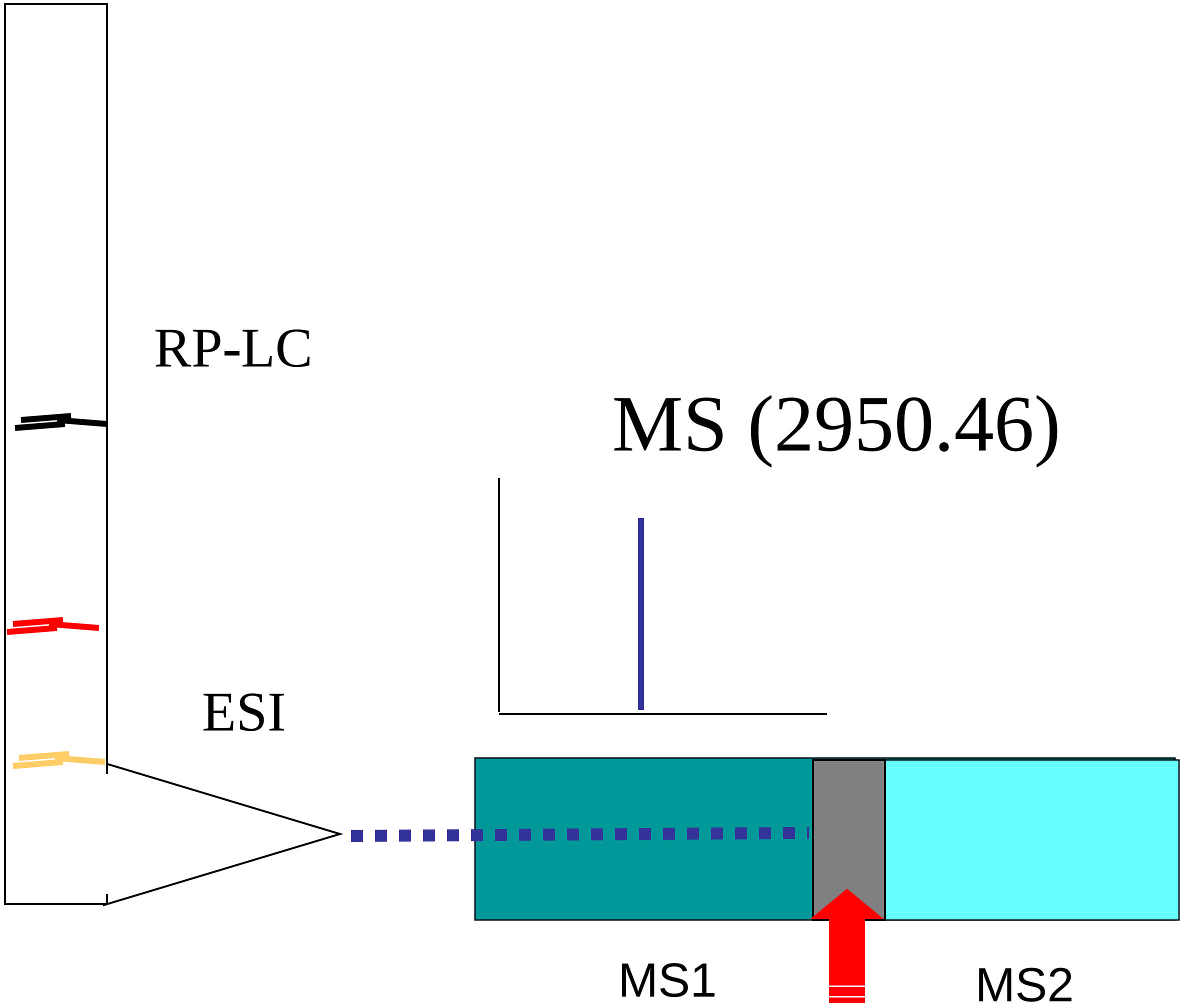
MS (2950.46)

ESI

MS1

MS2

LC-MS/MS



RP-LC

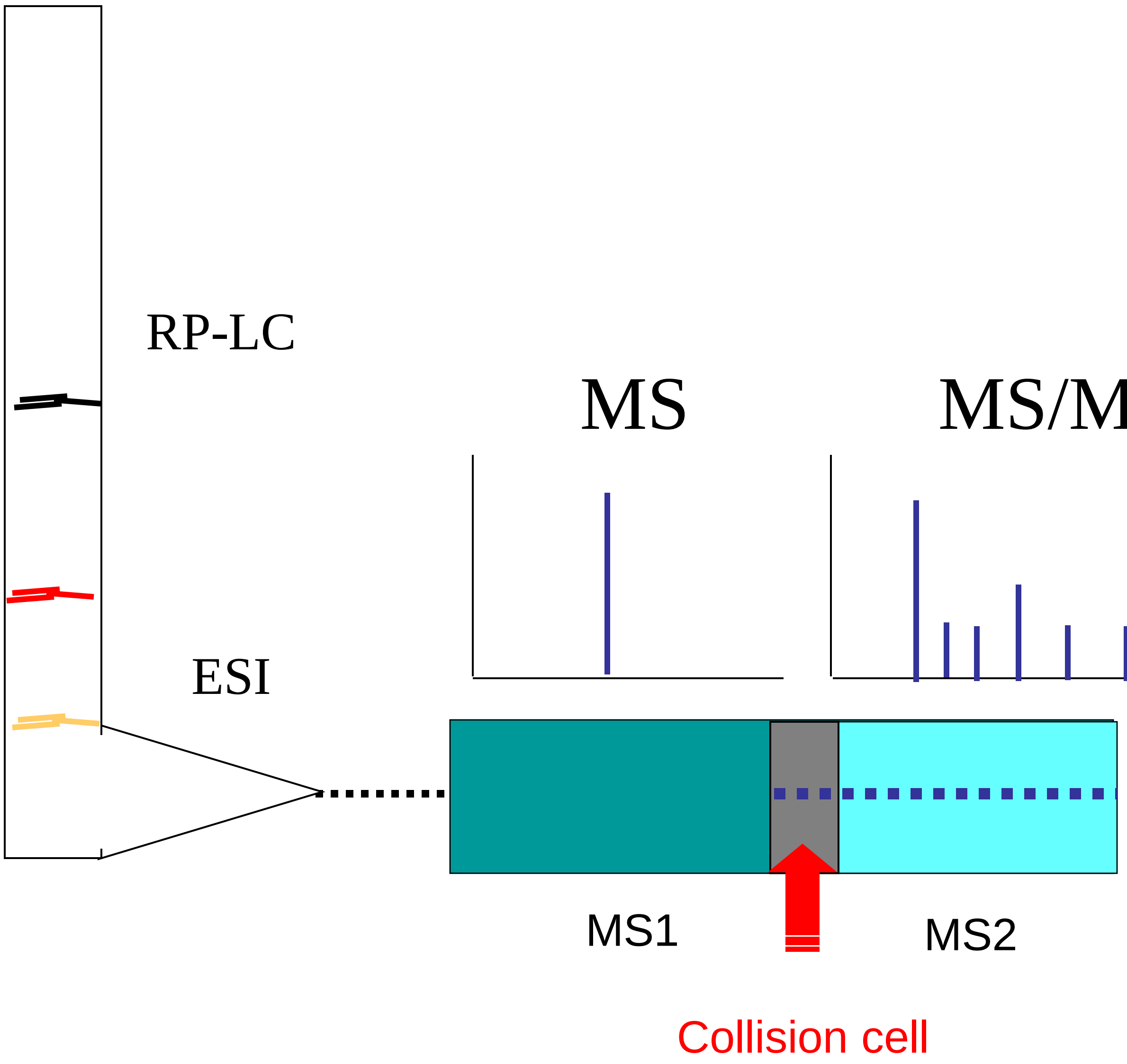
MS (2950.46)

ESI

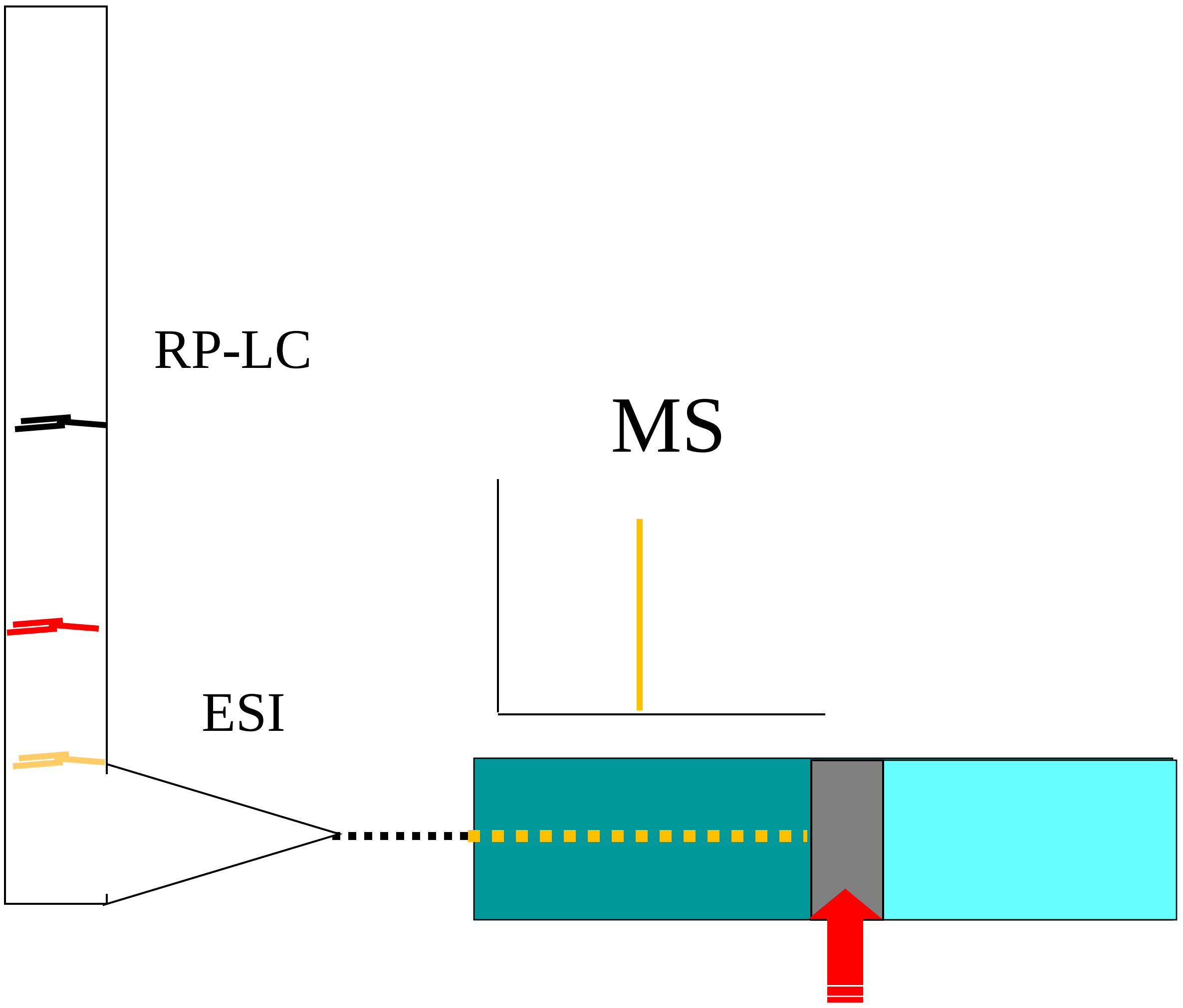
MS1

MS2

LC-MS/MS



LC-MS/MS

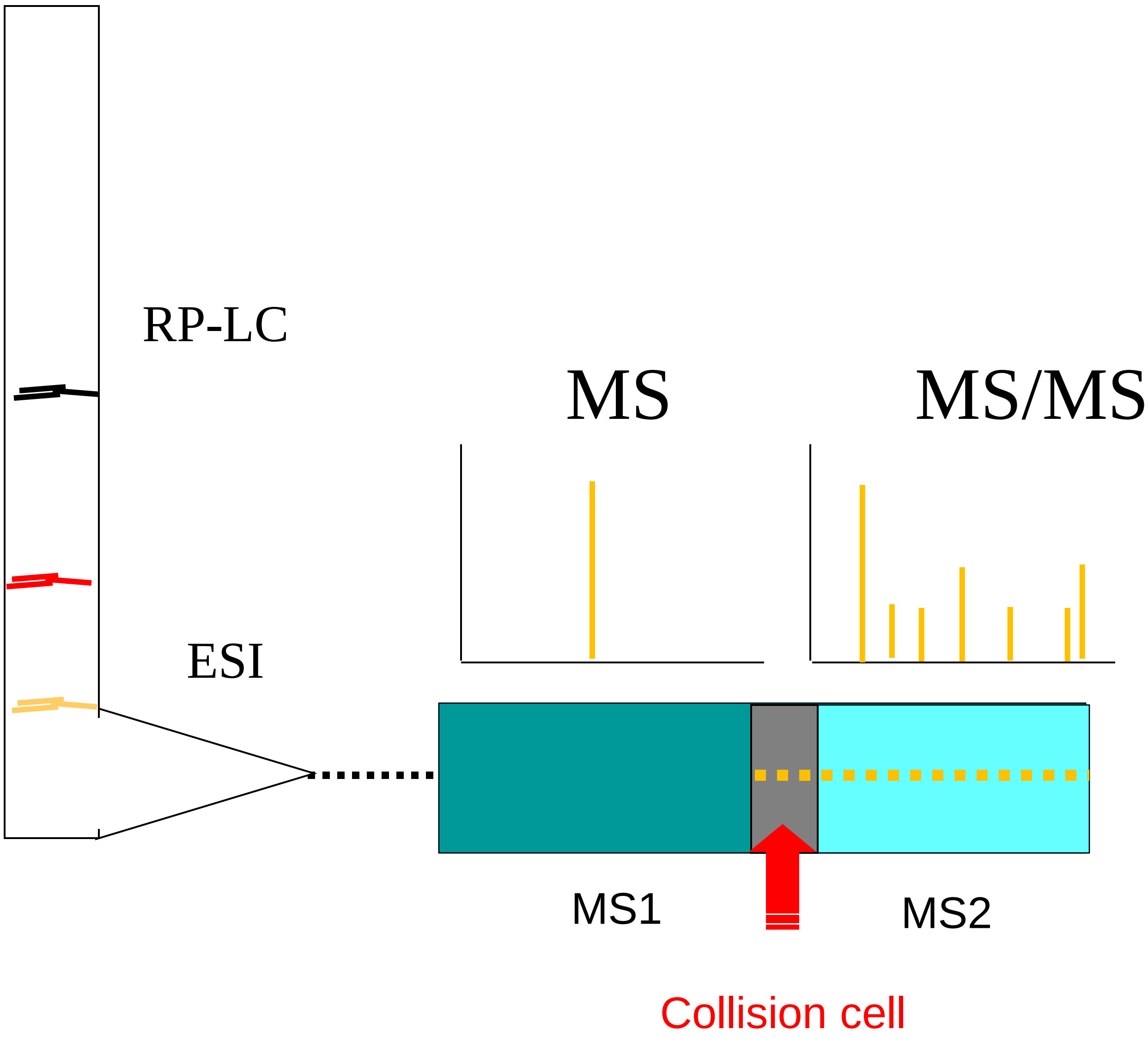


RP-LC

MS

ESI

LC-MS/MS



RP-LC

MS

MS/MS

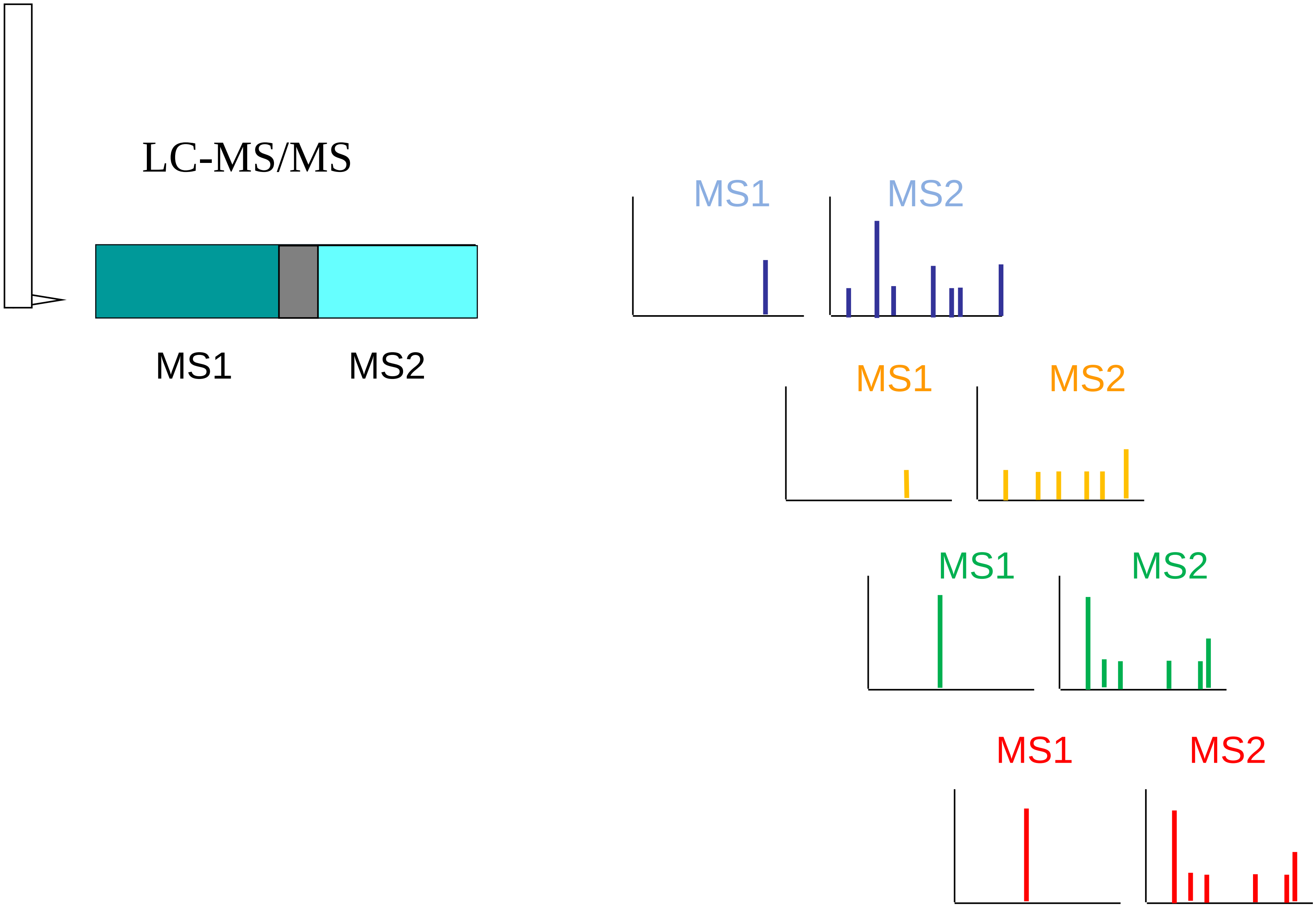
ESI

MS1

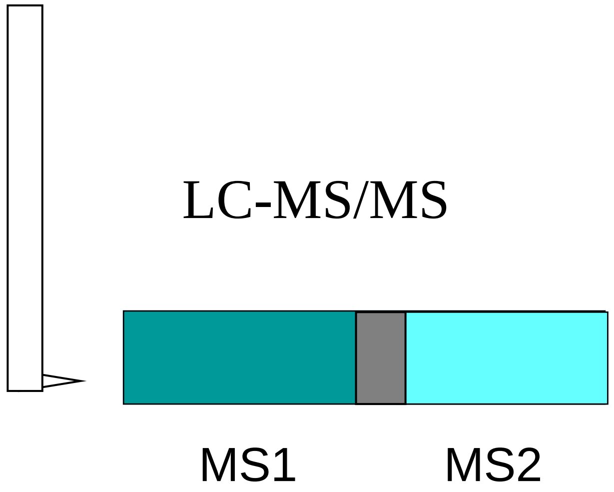
MS2

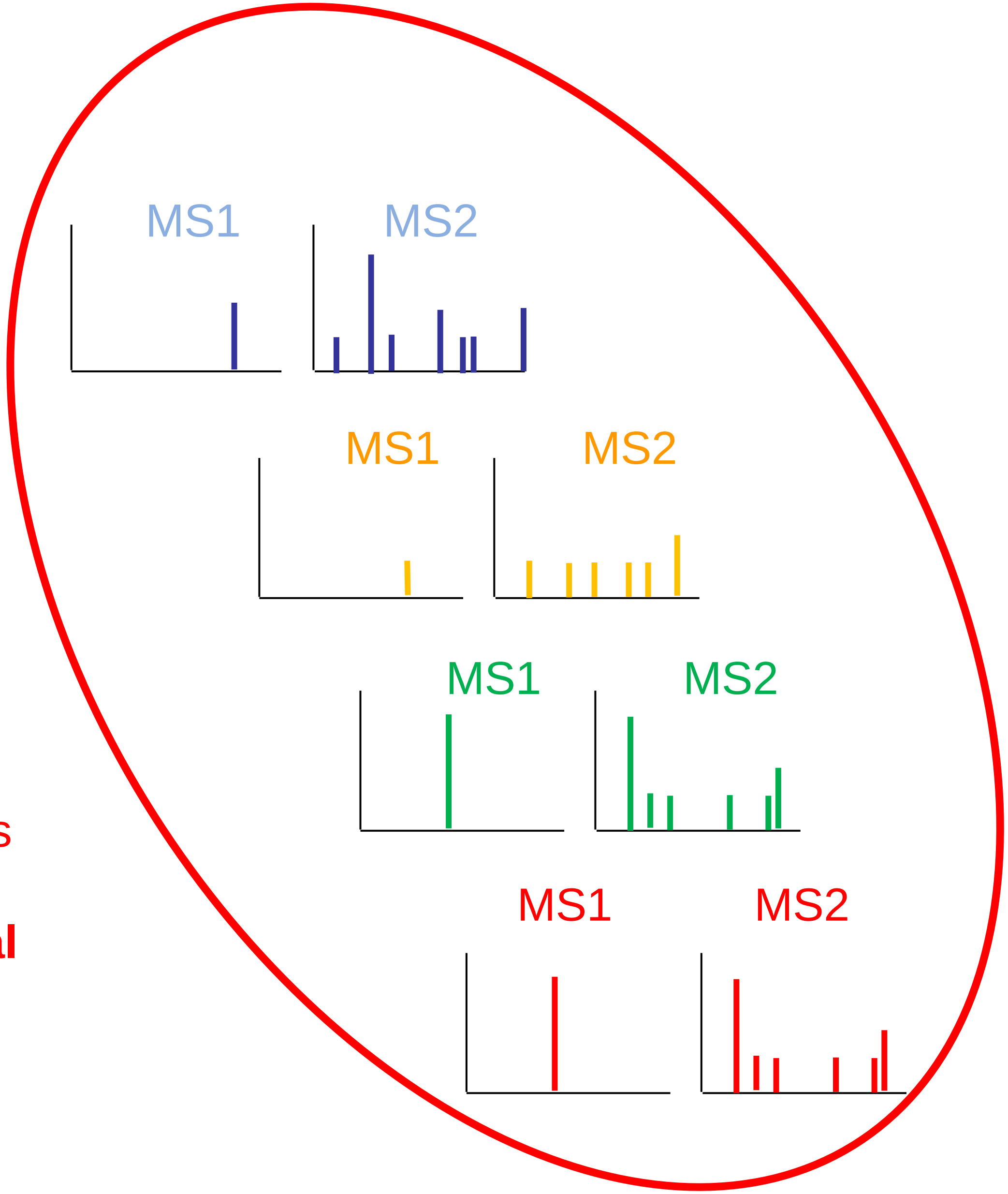
Collision cell

LC-MS/MS



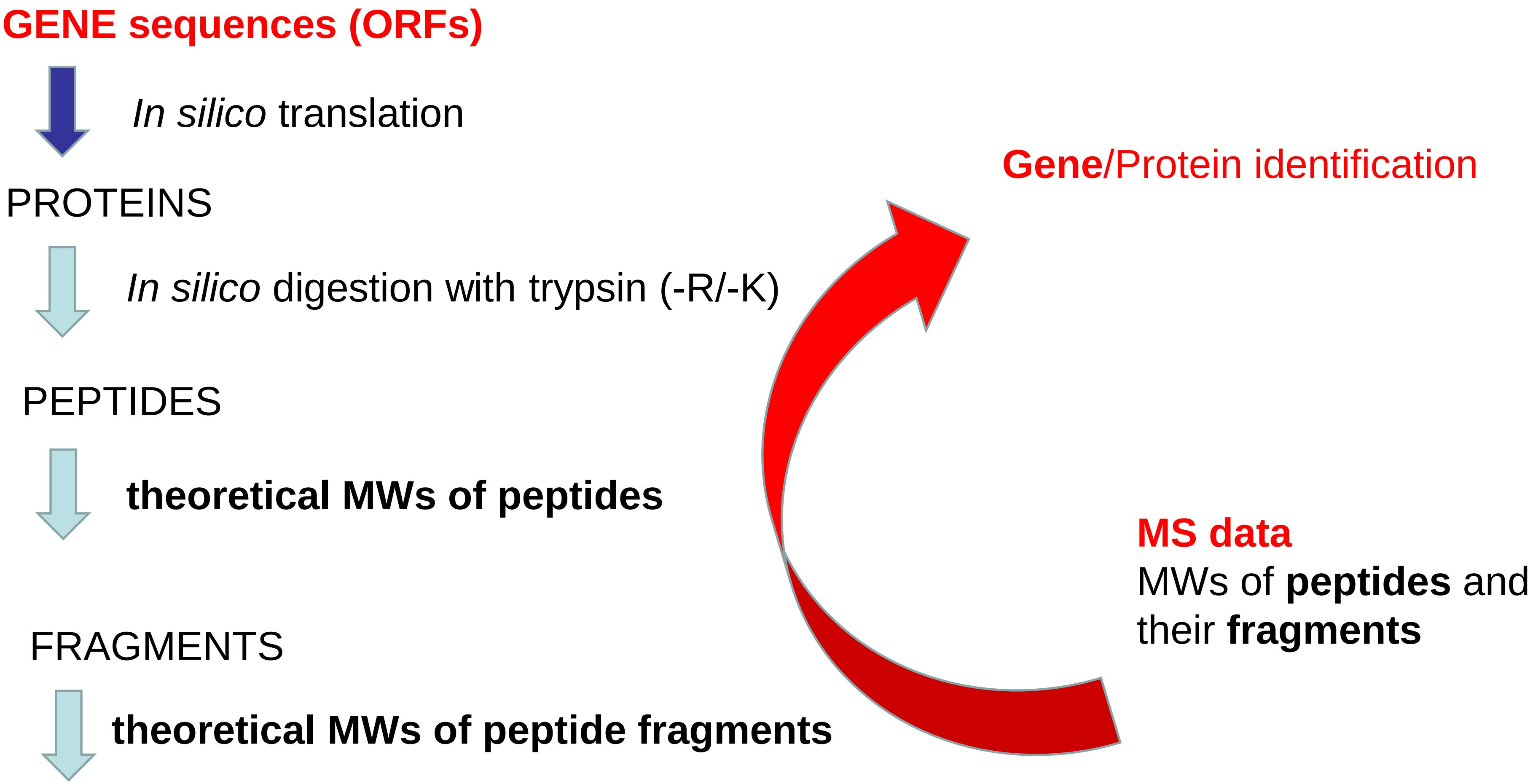
LC-MS/MS





The masses of the peptides and their fragments are **compared with theoretical masses** of all peptides and their fragments present in gene/protein databases

Identification of protein in databases based on MS/MS data



GENE sequences (ORFs)

In silico translation

PROTEINS

In silico digestion with trypsin (-R/-K)

PEPTIDES

theoretical MWs of peptides

FRAGMENTS

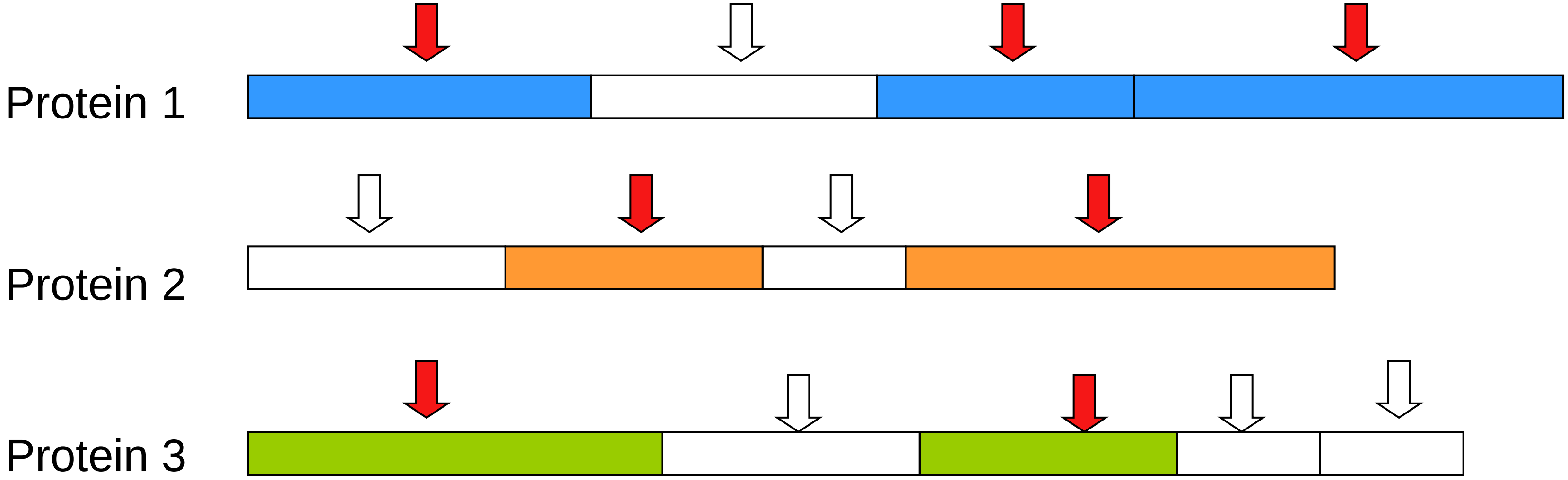
theoretical MWs of peptide fragments

Gene/Protein identification

MS data
MWs of **peptides** and
their **fragments**

Protein identification in shot-gun experiments

Unique peptides X peptides present in more than 1 protein



Keratin, type II cytoskeletal 4 (KRT4, P19013)

MIARQQCVRGGPRGFSCGSAIVGGGKRGAFSSVSMSGGAGRCSSGGFGSRSLYNL
RGNKSISMSVAGSRQGACFGGAGGFGTGGFGGGFGGSFSGKGGPGFPVCPAGGI
QEVTINQSLLTPLHVEIDPEIQKVRTEEREQIKLLNNKFASFIDKVQFLEQQNKVLETK
WNLLQQQTTTTSSKNLEPLFETYLSVLRKQLDTLGNDKGRLQSELKTMQDSVEDFK
TKYEEEINKRTAAENDFVVLKKDVDAAYLNKVELEAKVDSLNDEINFLKVLYDAELSQ
MQTHVSDTSVVLSMDNNRNLDLDSIIAEVRAQYEEIAQRSKAEAEALYQTKVQQL
QISVDQHGDNLKNTKSEIAELNRMIQRLRAEIENIKKQCQTLQVSVADAEQRGENA
LKDAHSKRVELEAALQQAKEELARMLREYQELMSVKLALDIEIATYRKLLEGEEYRM
SGECQSAVSISVVSGSTSTGGISGGLGSGSGFGLSSGFGSGSGSGFGFGGSVSGSSS
SKIISTTTLNKRR

Unique tryptic peptides (5-25 AA) represent only 38% in KRT4.

Search engines for MS/MS data processing/searching

MaxQuant (Andromeda) - *freely available*

MASCOT

SEQUEST

MSAmanda

Post-Translational Modifications

!!! Change in MW !!!

Phosphorylation (+80 Da)
Glycosylation (+162,203,146 Da)
S-S bonds (-2 Da)
Deamidation (+1 Da)
Gln -> pyro-Glu (-17 Da)
Oxidation (+16 Da)
Acetylation (+42 Da)
Acylation (+236/238 Da, etc.)
Loss of N-term Met (-131 Da)
ADP ribosylation (+541 Da)

Courtesy of Petr Man

Proteomic workflow
LC-MS/MS

proteins

trypsin…

peptides

1D or 2D separation

ON-LINE

Mass spectrometry

Peptide/protein identification

AND QUANTITATION

Quantitation of proteins/peptides in proteomics

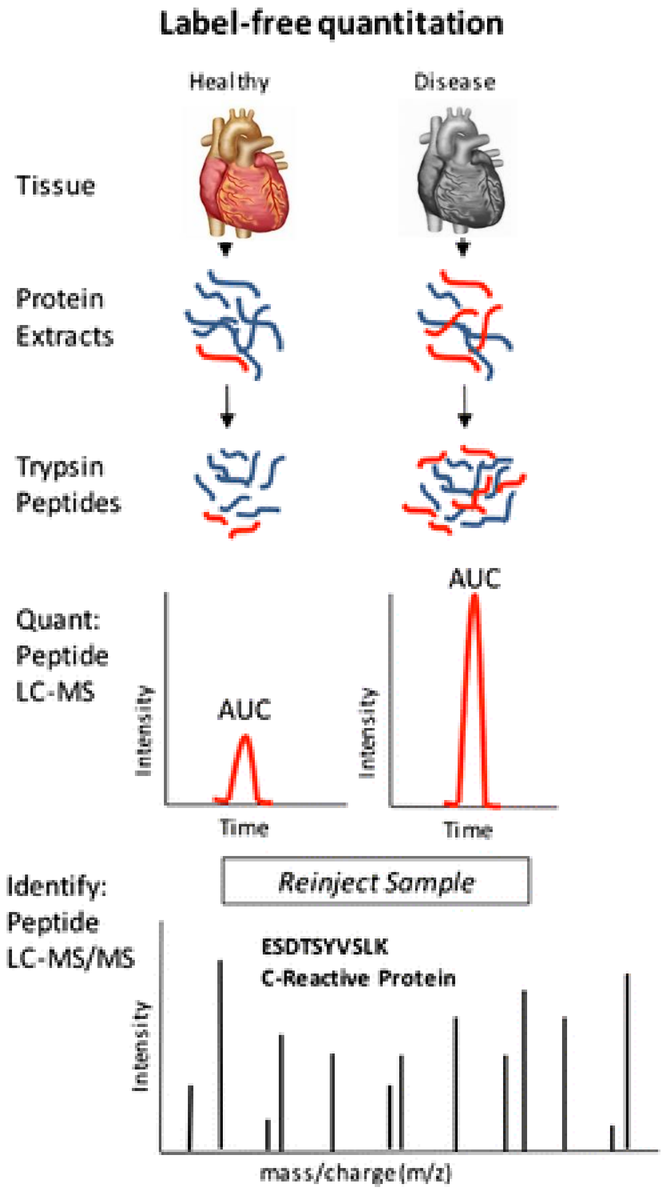
- Ionization and detection depend on the properties of the peptide
- Peptides compete with each other, one may suppress the signal of another

Direct quantitation of peptides/proteins by MS is problematic

Quantification of proteins/peptides in proteomics

1. **Label-free quantification**
2. **Quantification using stable-isotope labeling and multiplexing**

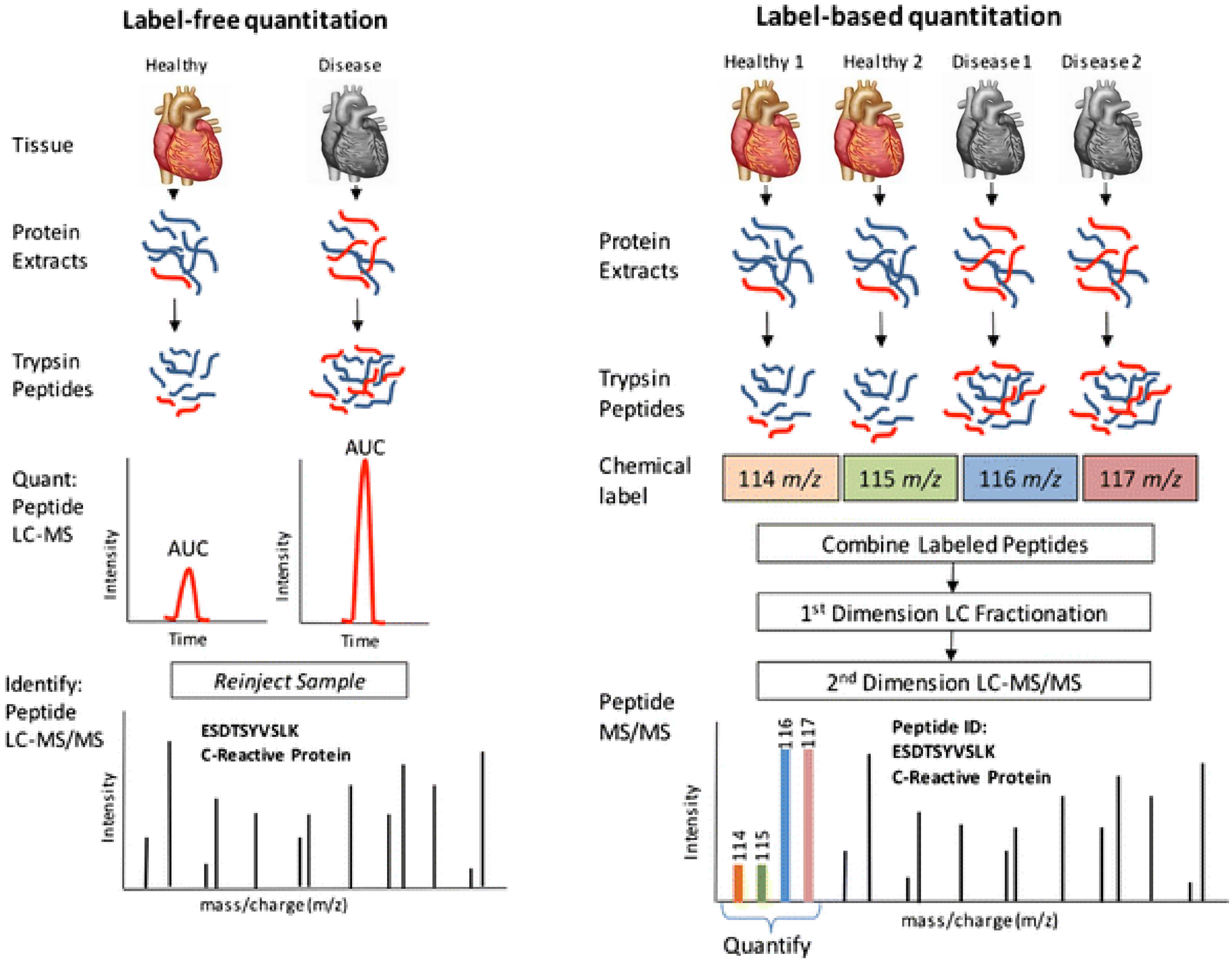
Protein quantitation in proteomics



Quantitation of proteins/peptides in proteomics

1. Label-free quantification
2. **Quantitation using stable-isotope labeling and multiplexing**

Protein quantitation in proteomics



Stable-isotope labeling and multiplexing

^{13}C, ^{15}N

- **Metabolic incorporation**
  metabolic incorporation of essential **amino acids containing SI** into proteins (live cells in cell culture, live animals)
  SILAC

- **Chemical incorporation**
  covalent modification of peptides with **tags containing SI**
  (any protein sample)
  TMT, iTRAQ

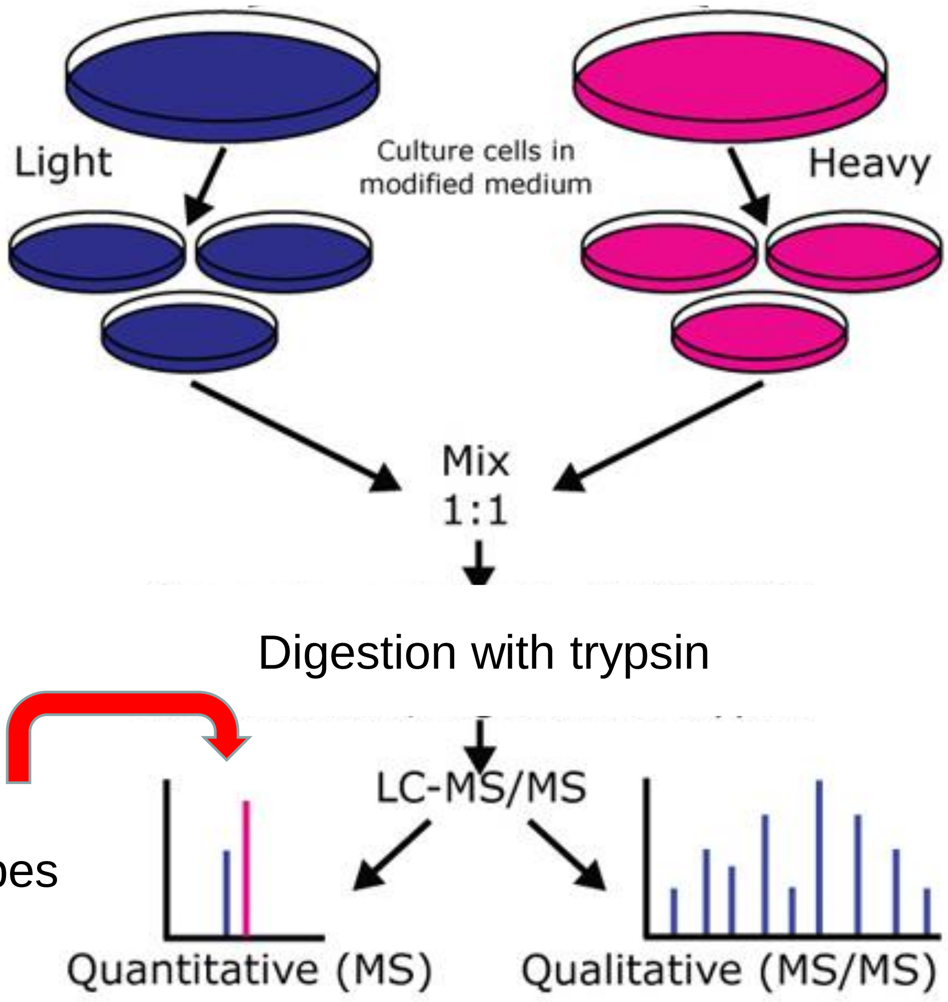
SILAC – Stable Isotope Labeling with Aminoacids in Culture

LIGHT

Normal AA
^{12}C

HEAVY

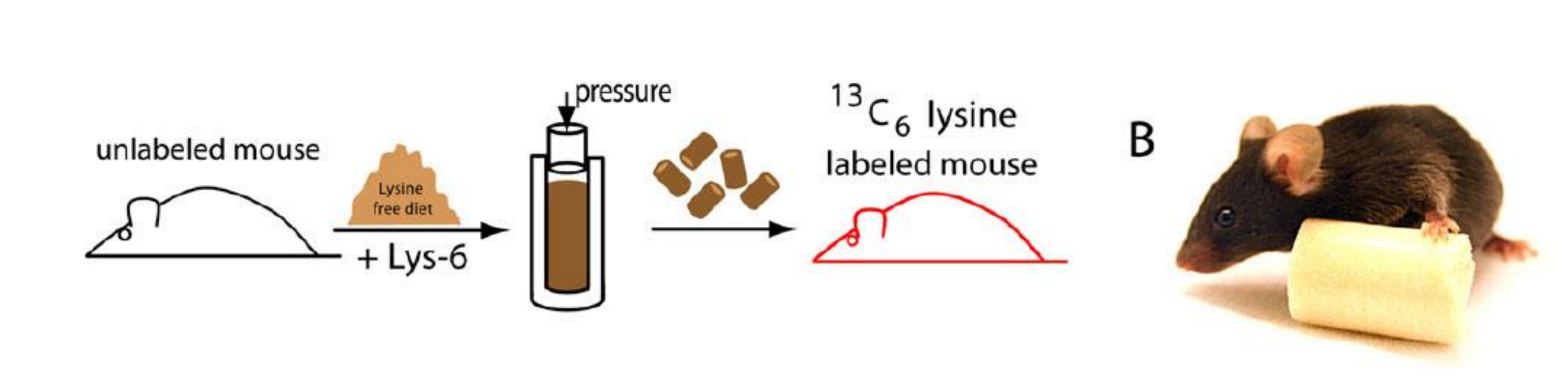
^{13}C-lysine
^{13}C-arginine



Digestion with trypsin

Defined mass shift
due to heavy isotopes

SILAC MICE

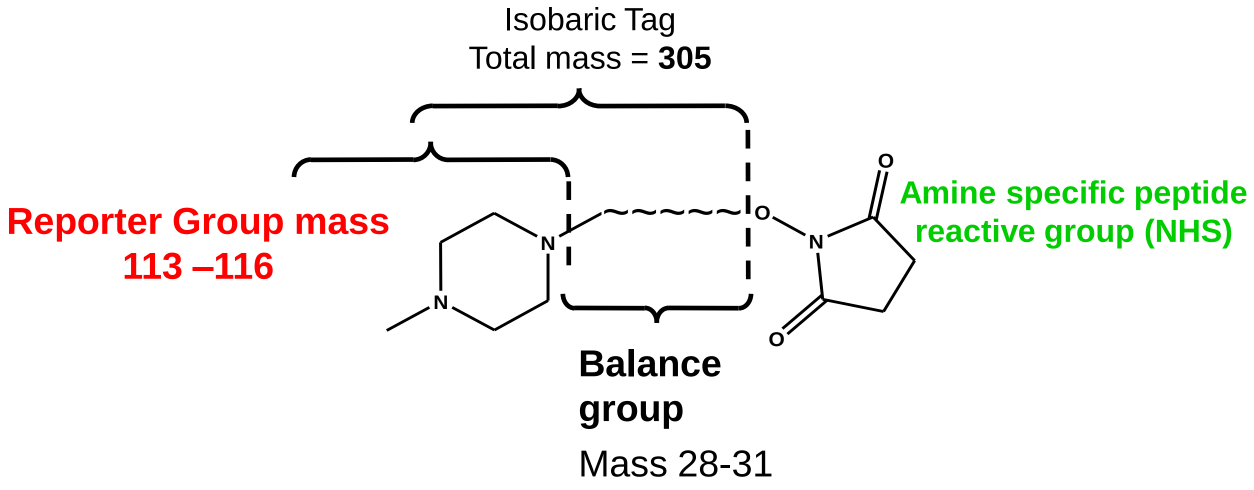


F1	93% $^{13}C_6$ lysine
F2	97% $^{13}C_6$ lysine
F3-F4	nearly 100 % $^{13}C_6$ lysine

Kruger, Moser et al. Cell *134*, 353–364

iTRAQ – isobaric Tags for Relative and Absolute Quantitation

N-termini and lysines in all peptides are labeled (after digestion with trypsin)



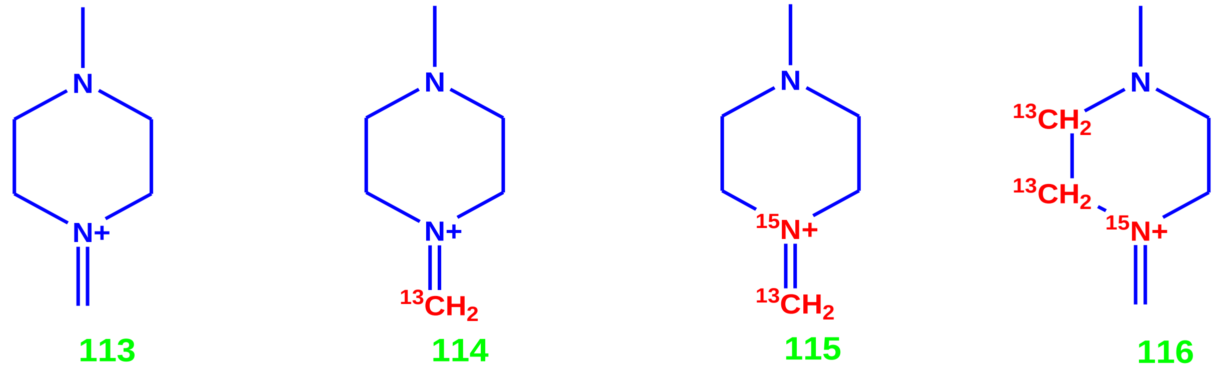
Isobaric Tag
Total mass = **305**

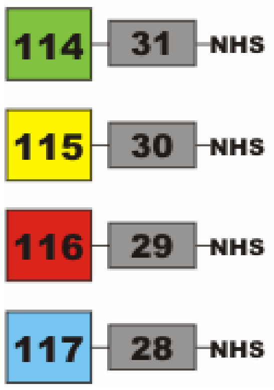
Reporter Group mass
113 –116

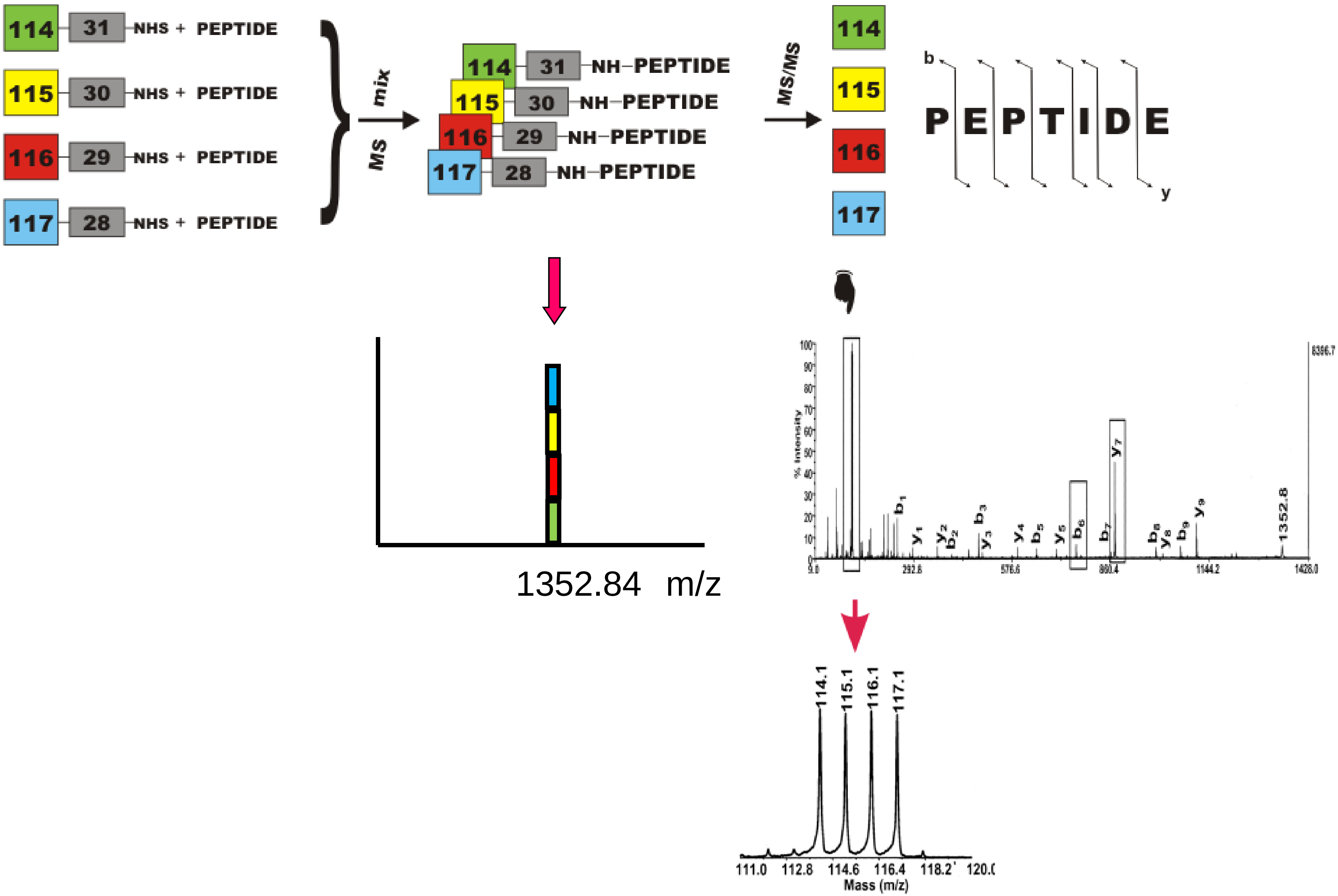
Balance group
Mass 28-31

Amine specific peptide reactive group (NHS)

Reporter Groups 113 –116



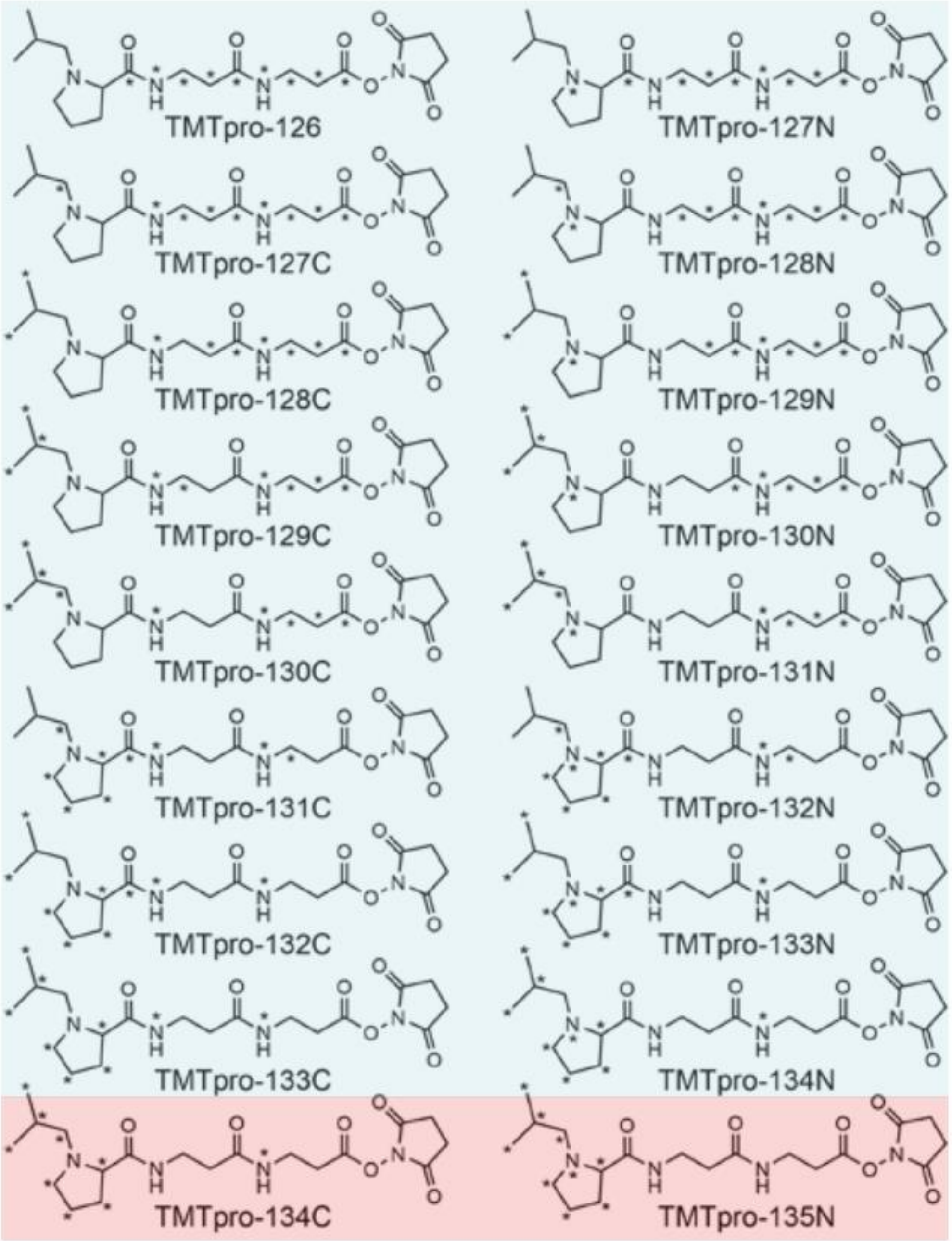



114 31 NHS + PEPTIDE
115 30 NHS + PEPTIDE
116 29 NHS + PEPTIDE
117 28 NHS + PEPTIDE
mix
MS
114 31 NH–PEPTIDE
115 30 NH–PEPTIDE
116 29 NH–PEPTIDE
117 28 NH–PEPTIDE
MS/MS
114
115
116
117
b
P E P T I D E
y
1352.84 m/z
% Intensity
b1 y1 y2 b2 b3 y3 y4 b5 y5 b6 b7 y7 b8 y8 b9 y9 1352.8
9.0 292.8 576.6 860.4 1144.2 1428.0
114.1 115.1 116.1 117.1
111.0 112.8 114.6 116.4 118.2 120.0
Mass (m/z)

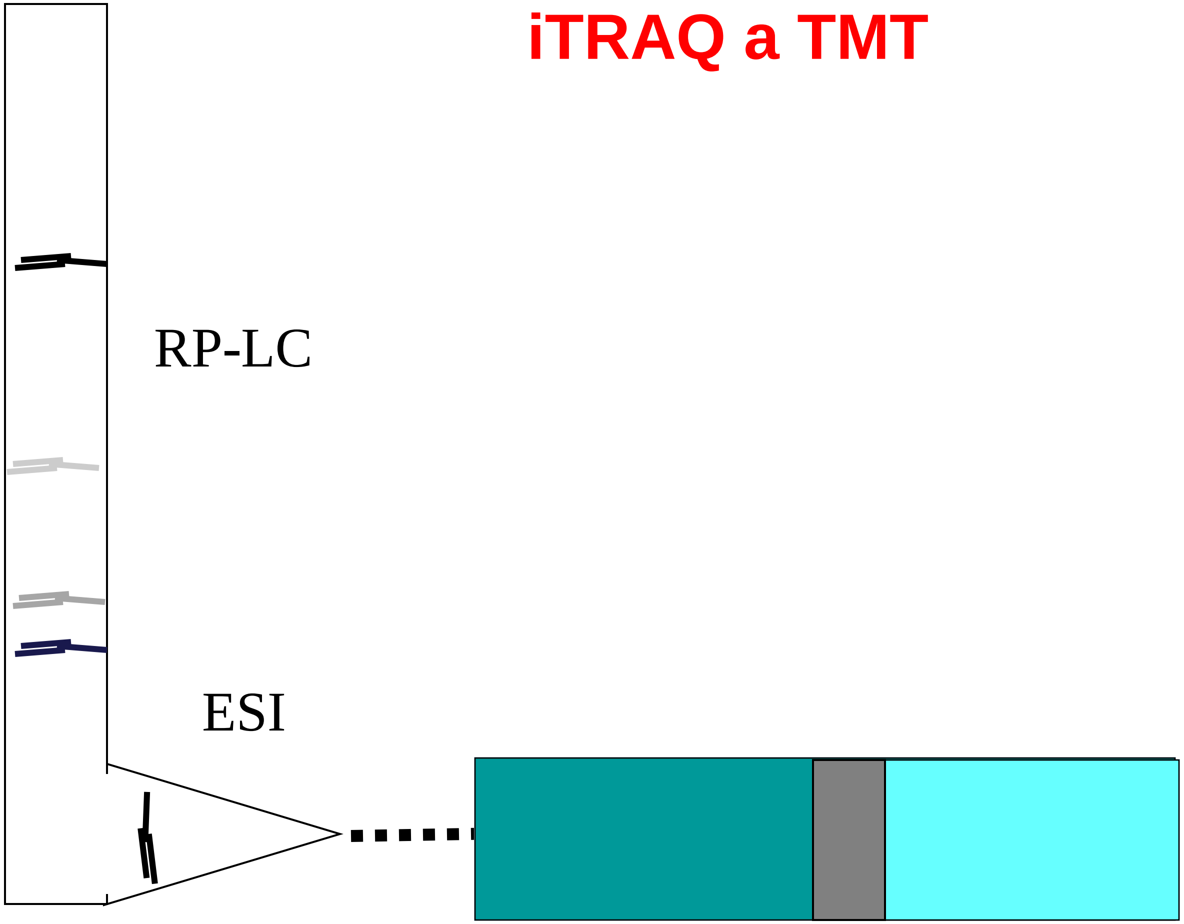
Isobaric tags TMT

(ThermoScientific)

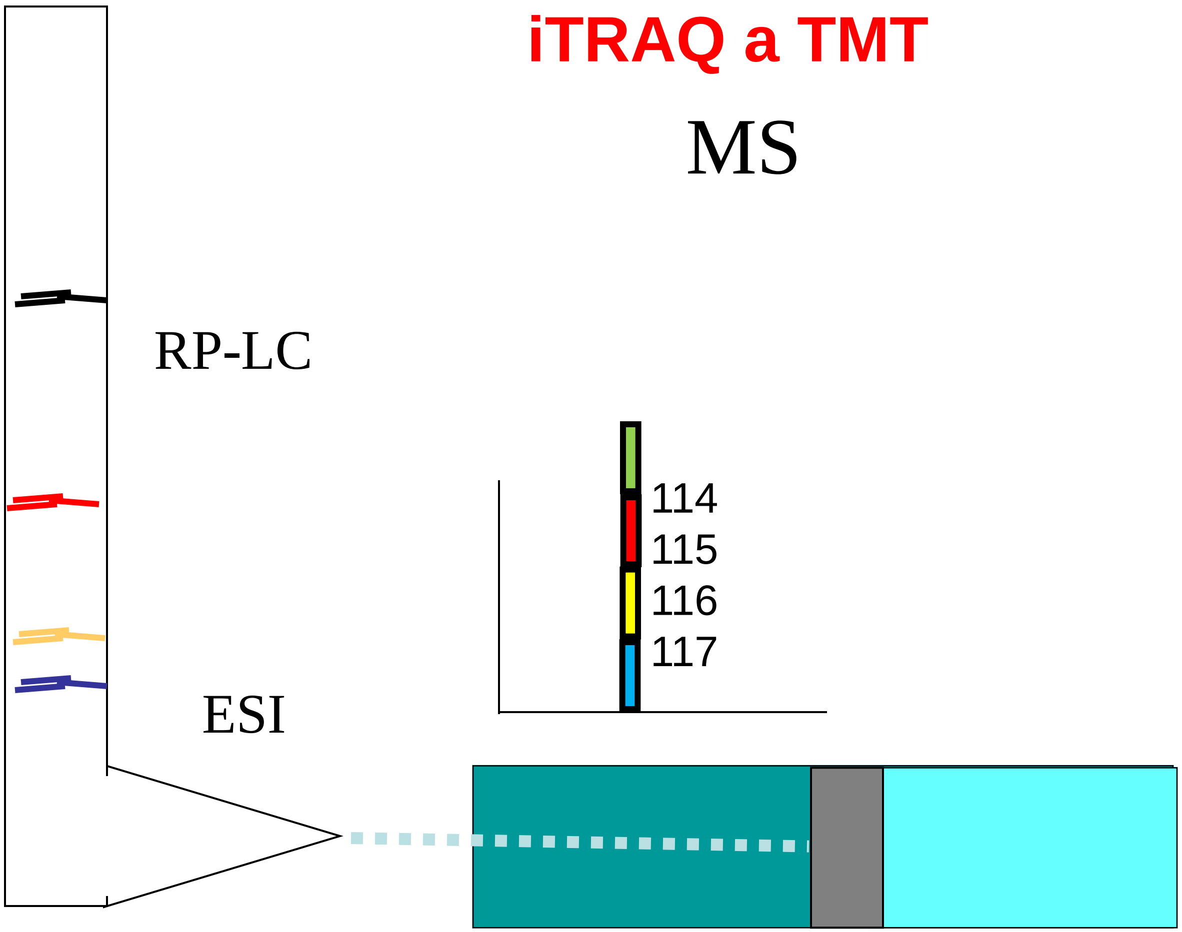
Up to 18 different mass tags



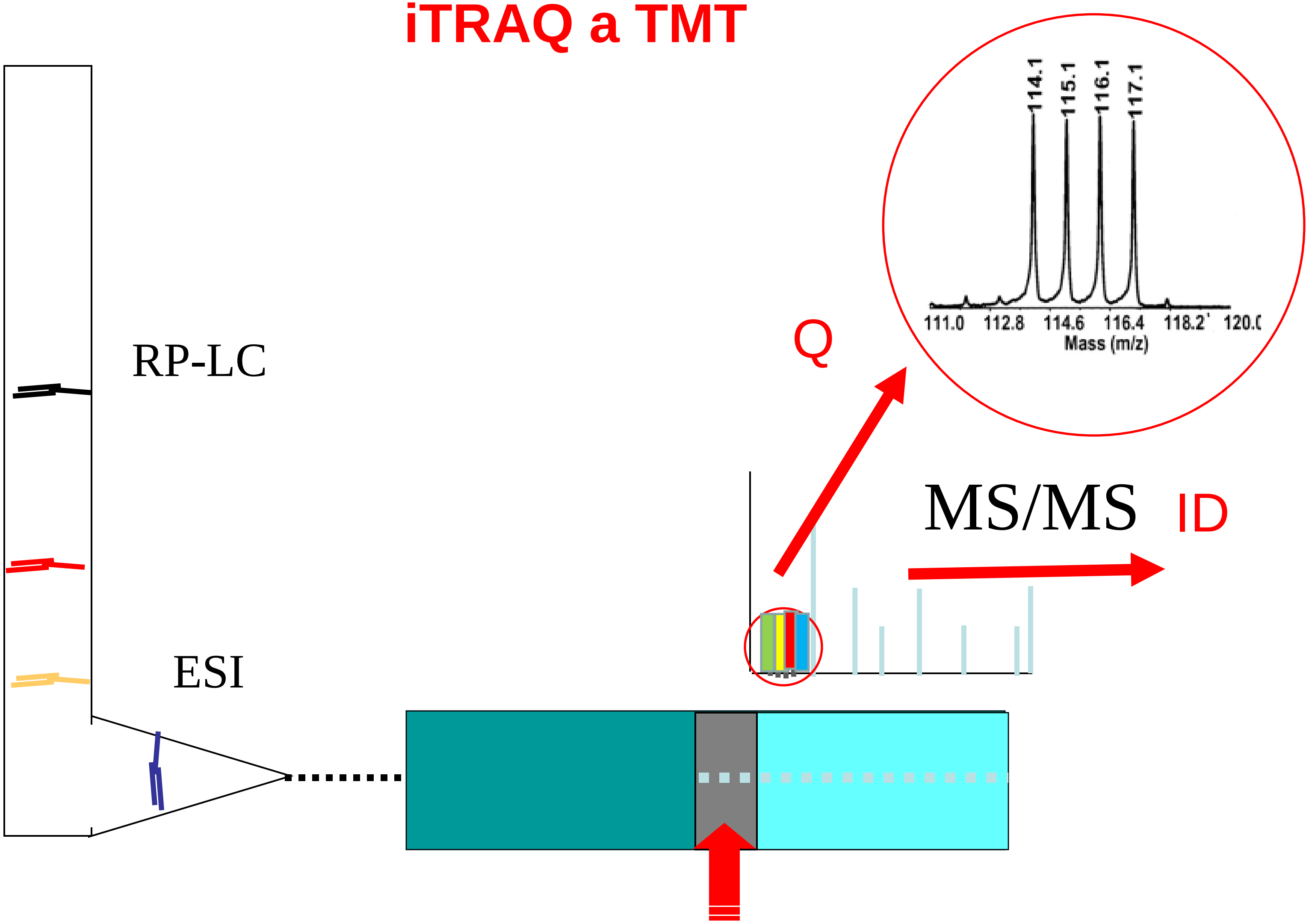
Isobaric labeling and multiplexing iTRAQ a TMT



Isobaric labeling and multiplexing
iTRAQ a TMT



Isobaric labeling and multiplexing iTRAQ a TMT



RP-LC

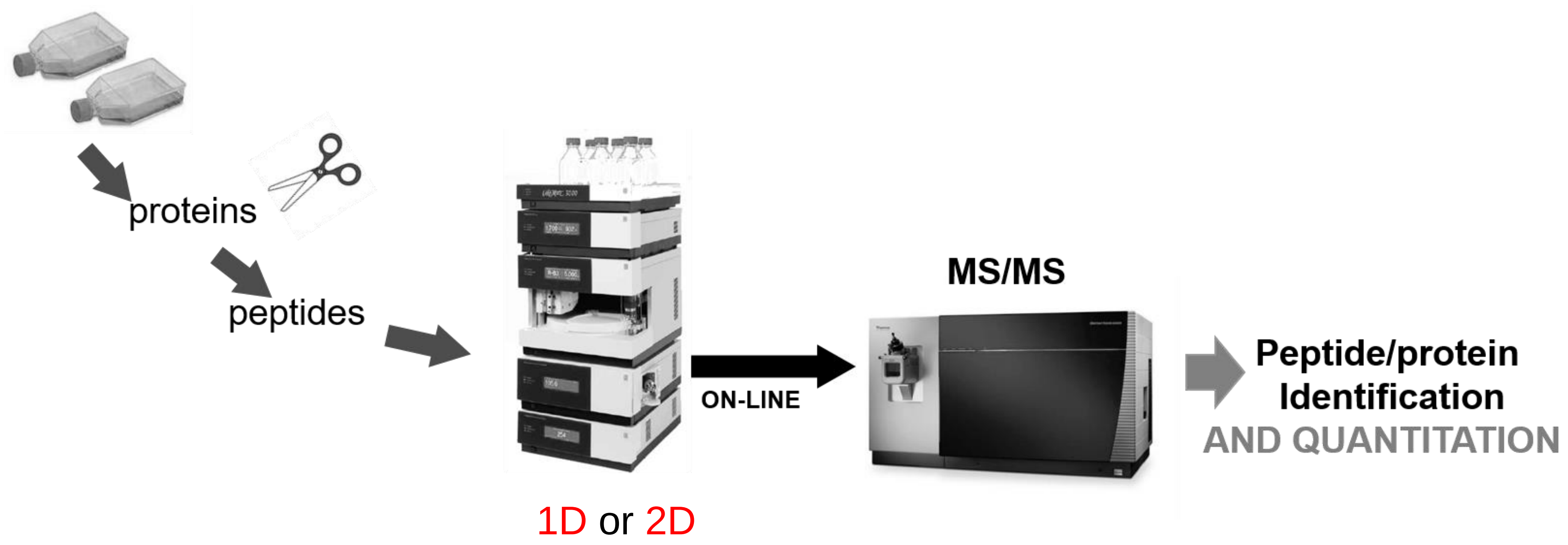
ESI

Q

MS/MS

ID

Current expression proteomics



proteins

peptides

ON-LINE

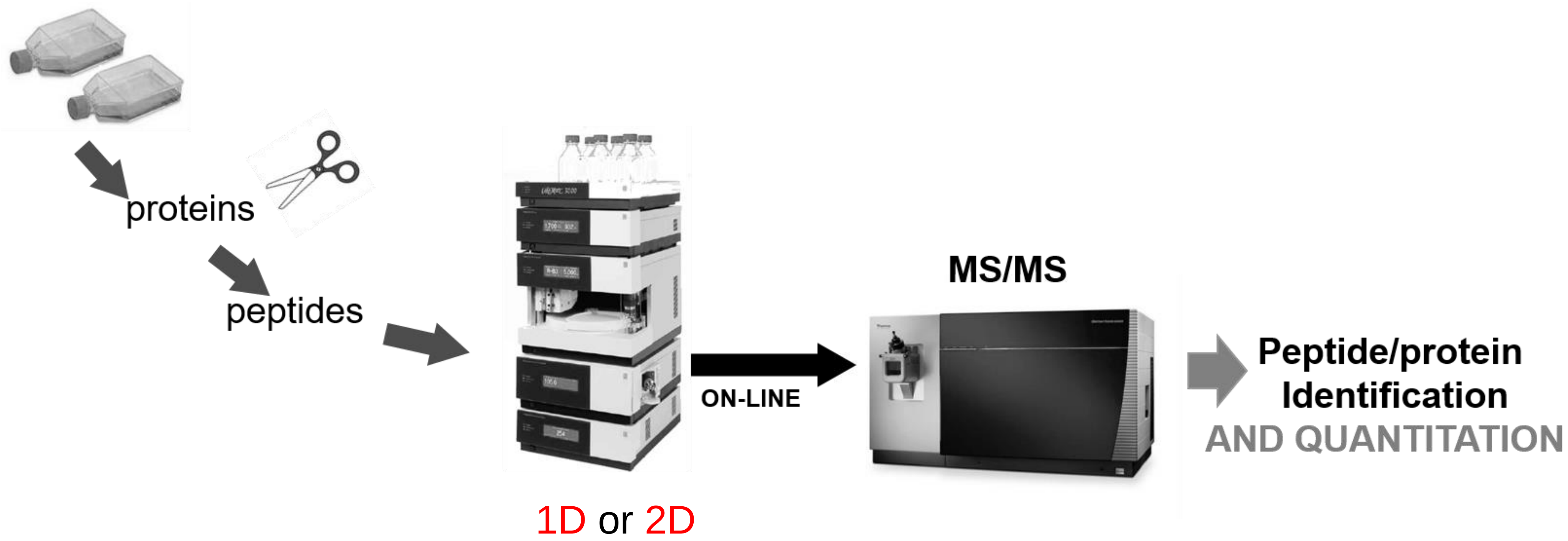
MS/MS

Peptide/protein Identification

AND QUANTITATION

1D or 2D

Current expression proteomics



proteins

peptides

ON-LINE

MS/MS

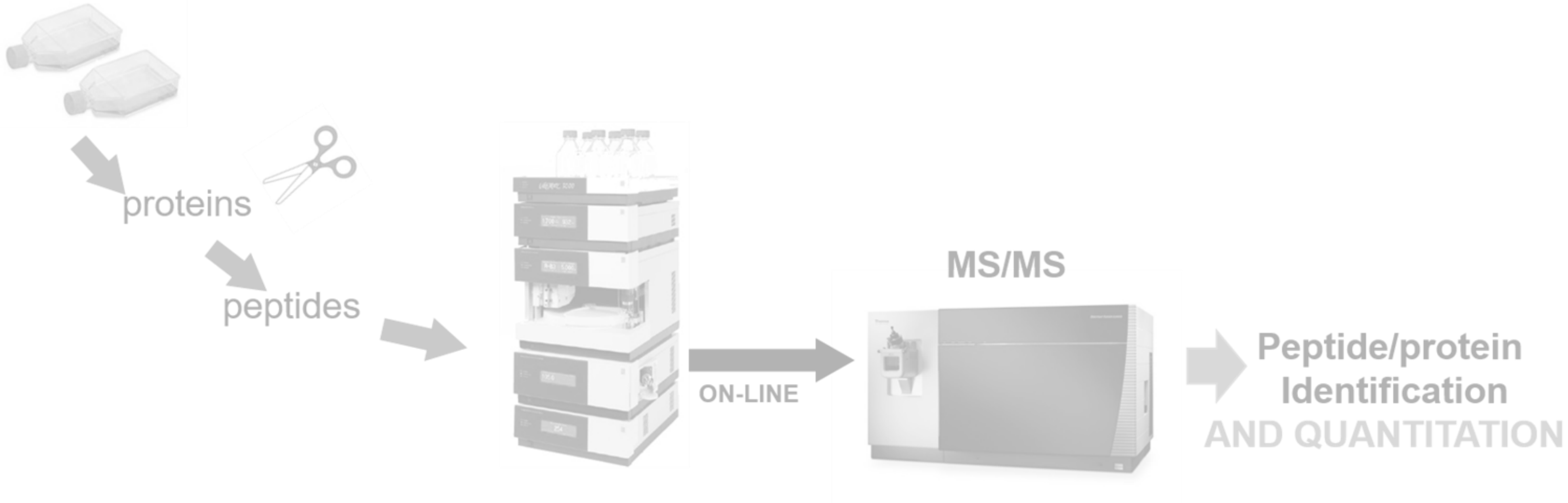
Peptide/protein Identification

AND QUANTITATION

1D or 2D

Human cellular tissues	**~ 5000 - 10 000**	**proteins**
Plasma/Serum	**~ 400 - 2000**	**proteins**
CSF	**~ 400 - 1500**	**proteins**
Urine	**~ 200 - 800**	**proteins**

A single cell	**over 4000**	**proteins**



proteins

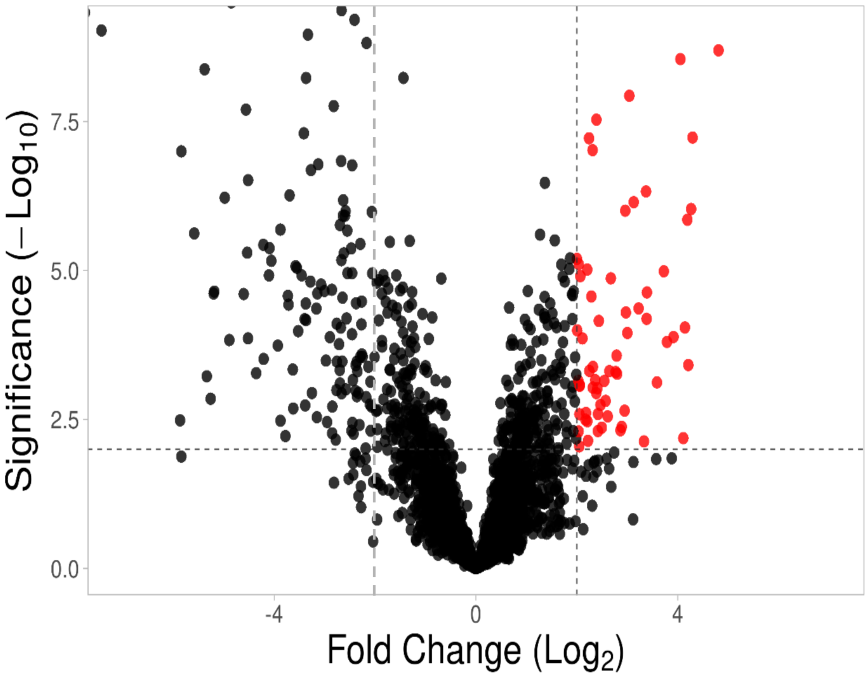
peptides

ON-LINE

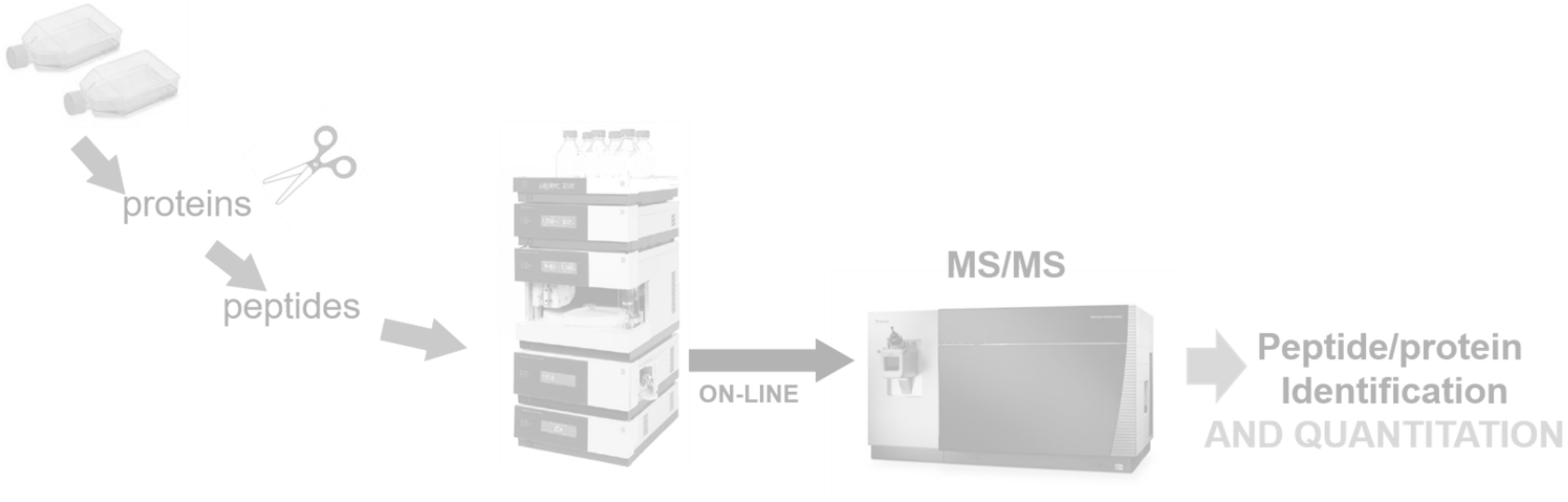
MS/MS

Peptide/protein Identification

AND QUANTITATION



Volcano plot



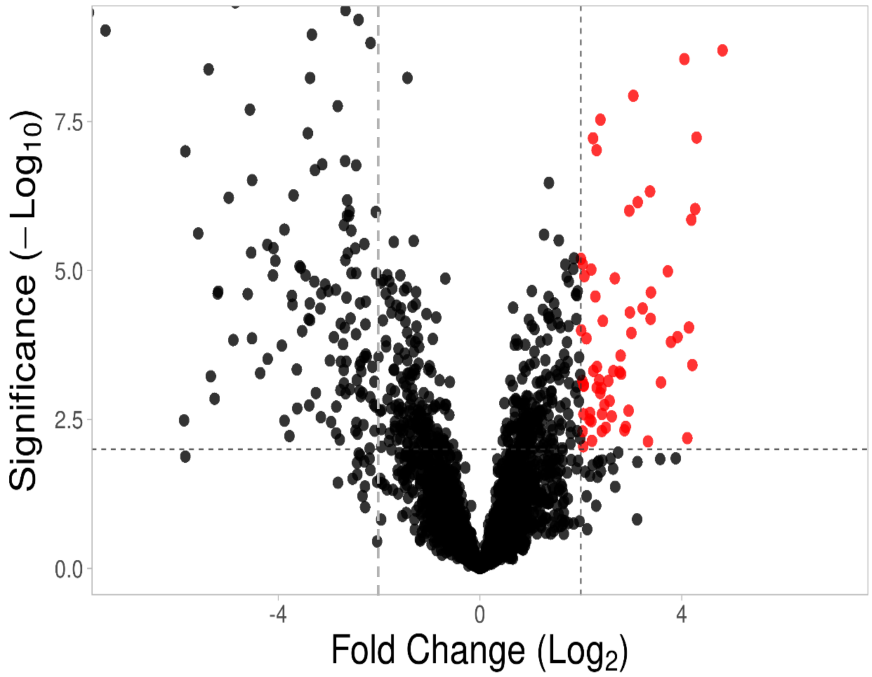
proteins

peptides

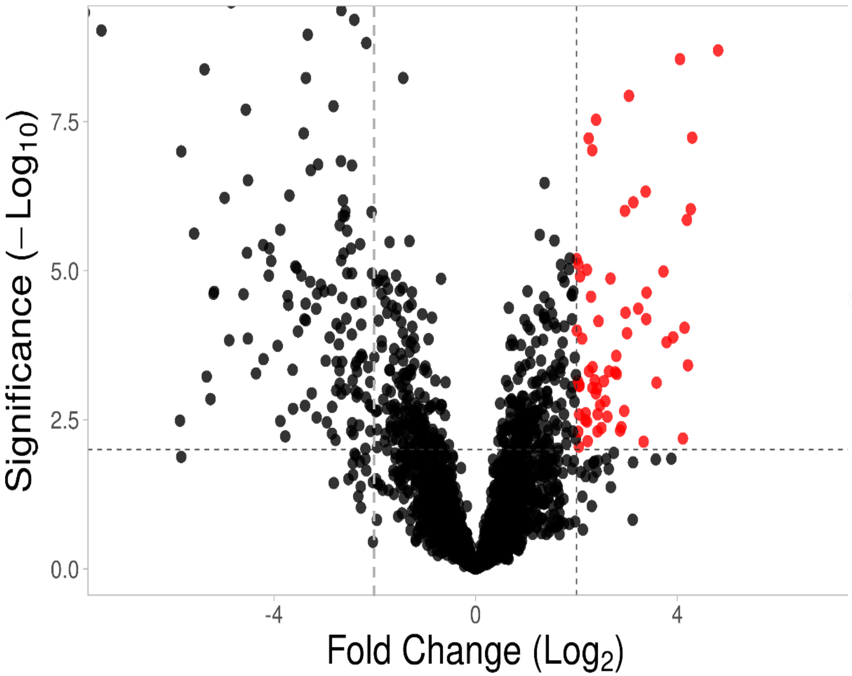
ON-LINE

MS/MS

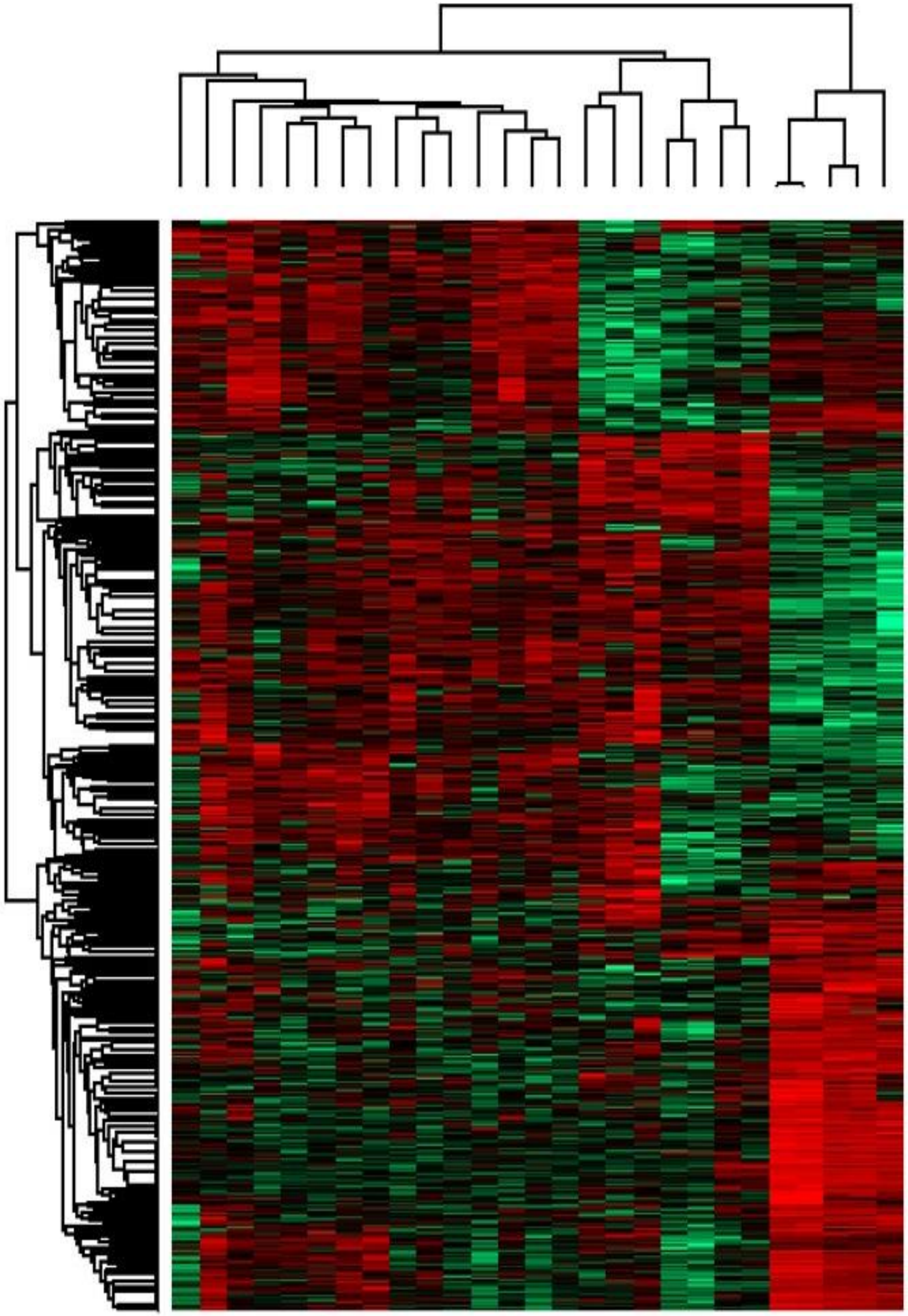
Peptide/protein Identification AND QUANTITATION



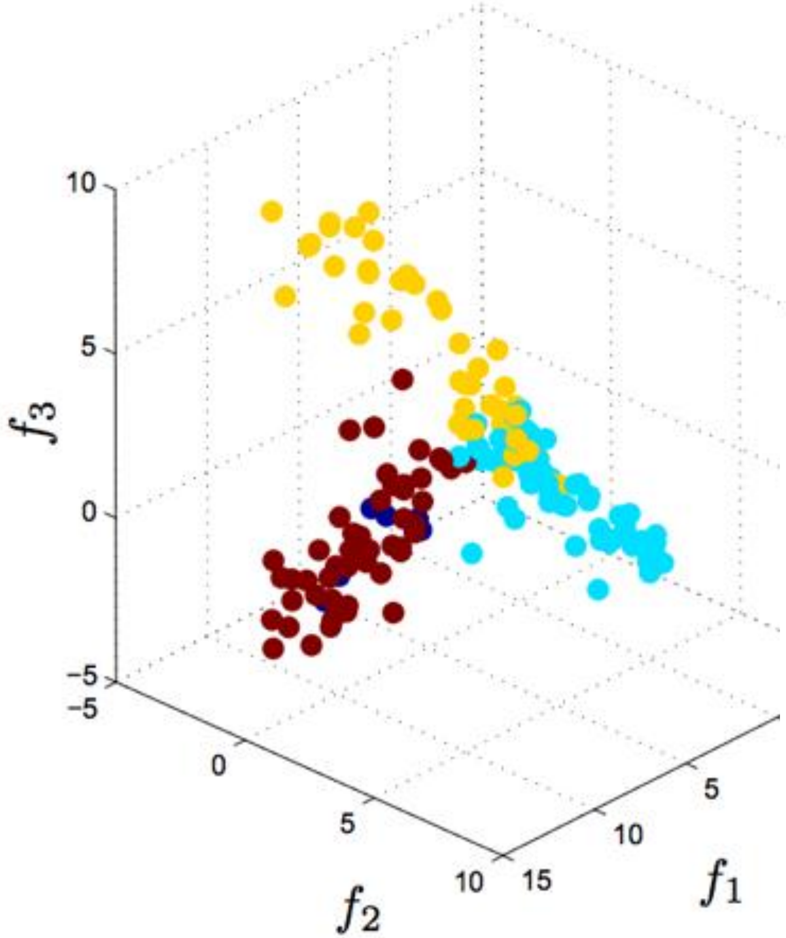
Volcano plot



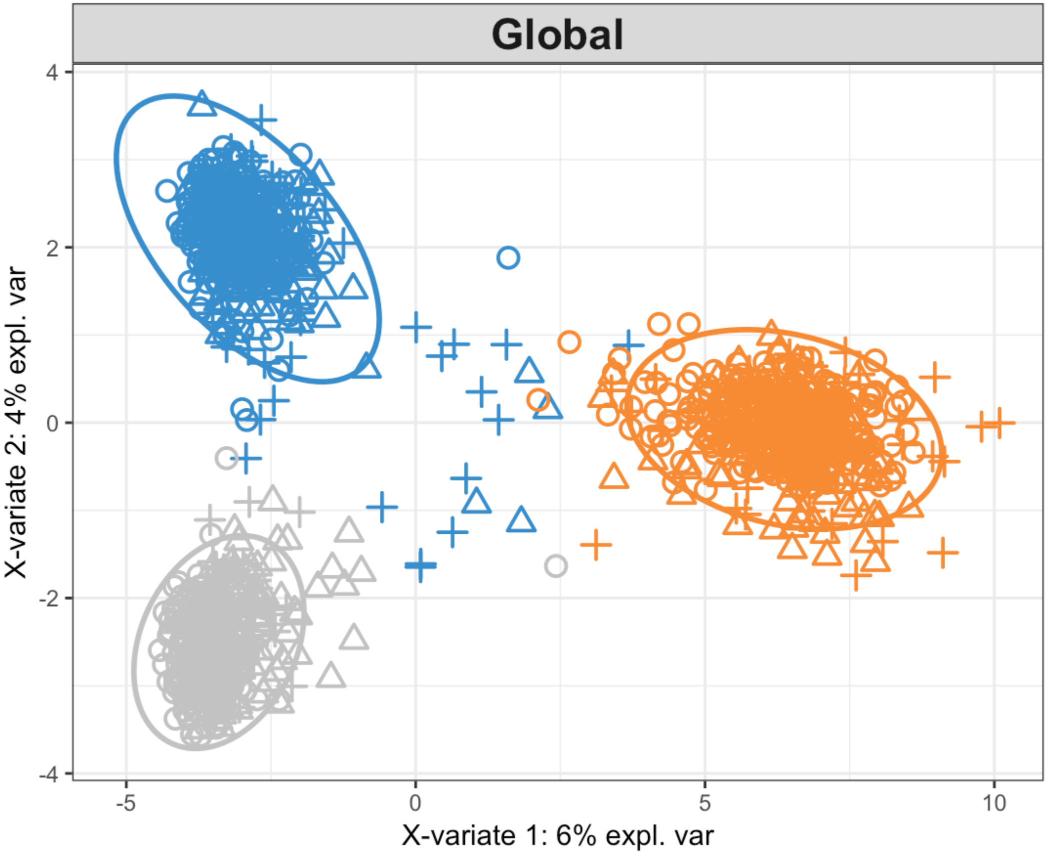
Volcano plot

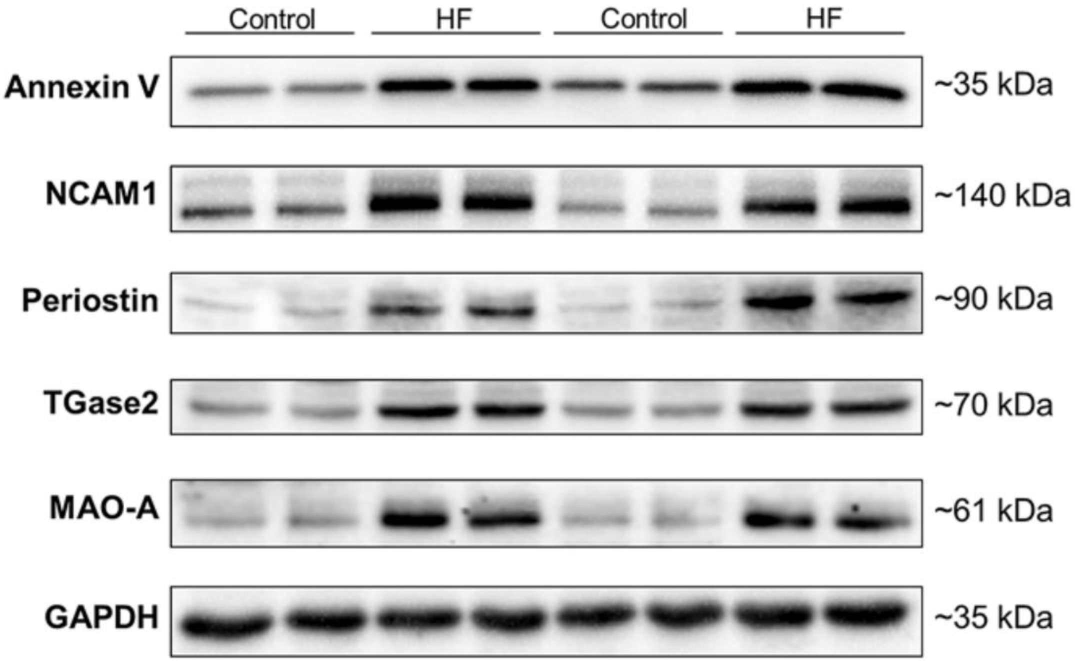


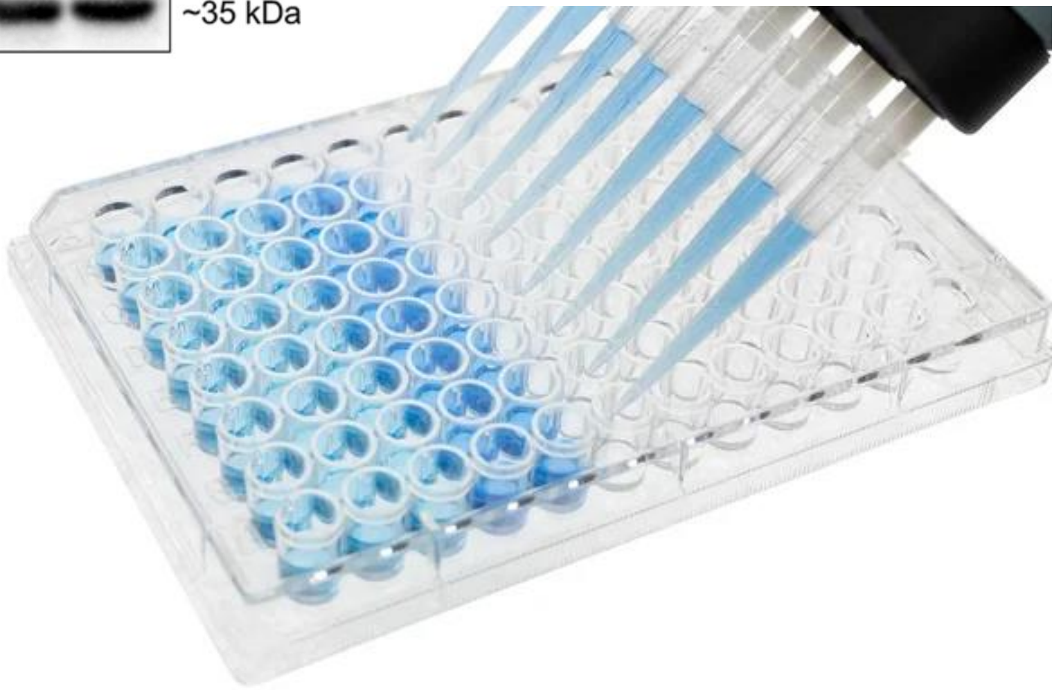
PCA

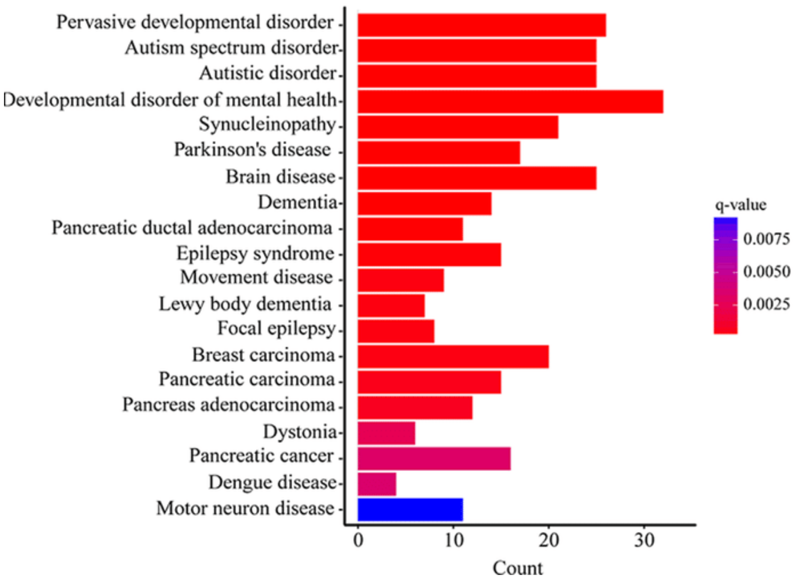


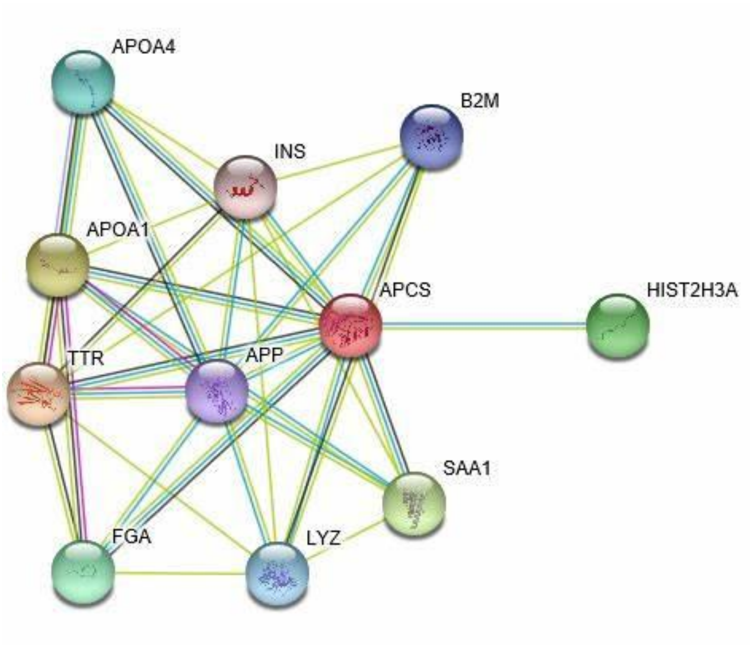
MINT sPLS-DA

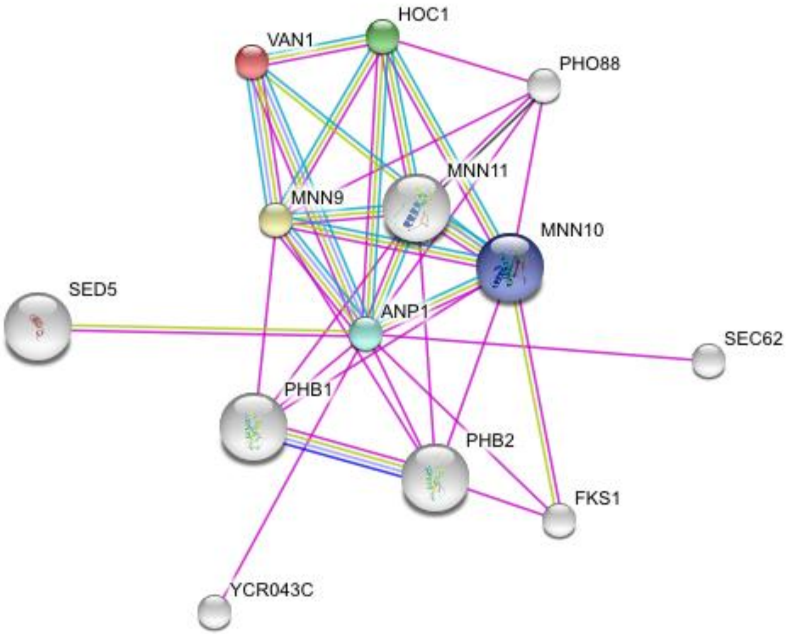



Control
HF
Control
HF
Annexin V
~35 kDa
NCAM1
~140 kDa
Periostin
~90 kDa
TGase2
~70 kDa
MAO-A
~61 kDa
GAPDH
~35 kDa




Pervasive developmental disorder
Autism spectrum disorder
Autistic disorder
Developmental disorder of mental health
Synucleinopathy
Parkinson's disease
Brain disease
Dementia
Pancreatic ductal adenocarcinoma
Epilepsy syndrome
Movement disease
Lewy body dementia
Focal epilepsy
Breast carcinoma
Pancreatic carcinoma
Pancreas adenocarcinoma
Dystonia
Pancreatic cancer
Dengue disease
Motor neuron disease
0
10
20
30
Count
q-value
0.0075
0.0050
0.0025


APOA4
B2M
INS
APOA1
APCS
HIST2H3A
TTR
APP
SAA1
FGA
LYZ


HOC1
VAN1
PHO88
MNN11
MNN9
MNN10
SED5
ANP1
SEC62
PHB1
PHB2
FKS1
YCR043C

The wrap-up

- Up to 8000-13000 proteins can be identified/quantified in a human tissue samples
- We are getting closer to comprehensive coverage of human proteome
- MS-based „identification of a protein“ is actually identification of its coding gene
- We can not distinguish between individual protein variants (proteoforms)
- Only micrograms of the peptide sample are needed (mgs of tissue)
- Single-cell analysis is getting reasonable coverage (over 4000 IDs!)
- Specific methods for PTM analysis are available
- Information is semi-quantitative not absolute
- Some proteins remain under-represented in standard proteomic analyses