Advanced Mass-Spectrometry based Proteomics: quantification and post-translational modifications

February 9th, 2018

Manfredo Quadroni
Patrice Waridel
Roman Mylonas

WORKSHOP II SCHEDULE 2018

- **9.00 Course start**
- **9.05 Recapitulation of basic concepts in proteomics**
- **9:15 Statistical validation of protein identification: advanced concepts**
- **9:30 Post-translational modifications: general concepts and analytical specificities**
- **9:45 Phosphorylation analysis**
- **10:15 PTM exercise**
- **10:30 Coffee break**
- **11:00 Discussion of exercise results**
- **11:15 Other biological modifications, unexpected PTMs, artefacts**
- **12:00 Lunch break**
- **13:00 Introduction to quantitative proteomics, label-free and labeled quantitation**
- **14:00 Perseus exercise for quantification**
- **15:00 Exercise discussion**
- **15:30 Coffee break**
- **15:50 Statistics and validation in quantitation, publication guidelines**
- **16:00 Targeted quantification, DIA, conclusions**
- **16.25 Short break**
- **16:30 Test**
- **17:00 End**

Red : exercises

Today's goals

1) Give some knowledge on the mass spectrometry (MS) techniques used in proteomics to identify post-translational modifications in complex mixtures and quantify proteins

2) Practical analysis of datasets and evaluation of results

STEPS

1) Strategies for identification of PTMs with MS data : concepts and examples

- Mascot

2) Protein quantification with labeled and label-free techniques: concept and examples

- Perseus

Teachers:

Manfredo Quadroni (UNIL, PAFL)
Patrice Waridel (UNIL, PAFL)
Roman Mylonas (UNIL, SIB/PAFL)

Main pipeline (bottom-up proteomics)

Sample preparation

Protein mixture

1.Digestion

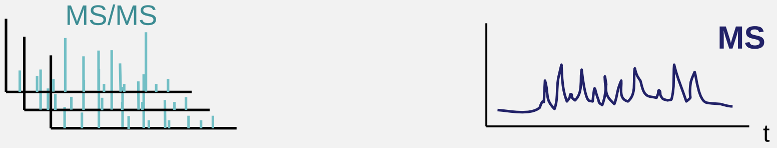
(trypsin : xxR.xx, xxK.xx)

+ reduction of disulfide bridges & alkylation of cysteines

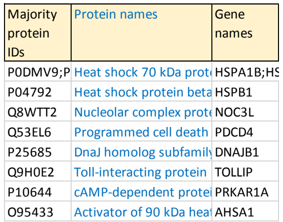
Peptide mixture

2. LC-MS/MS

MS Data



3. Database search



Majority protein IDs	Protein names	Gene names
P0DMV9;P	Heat shock 70 kDa prot	HSPA1B;HS
P04792	Heat shock protein beta	HSPB1
Q8WTT2	Nucleolar complex prot	NOC3L
Q53EL6	Programmed cell death	PDCD4
P25685	DnaJ homolog subfamily	DNAJB1
Q9H0E2	Toll-interacting protein	TOLLIP
P10644	cAMP-dependent protei	PRKAR1A
O95433	Activator of 90 kDa heat	AHSA1

Protein ID

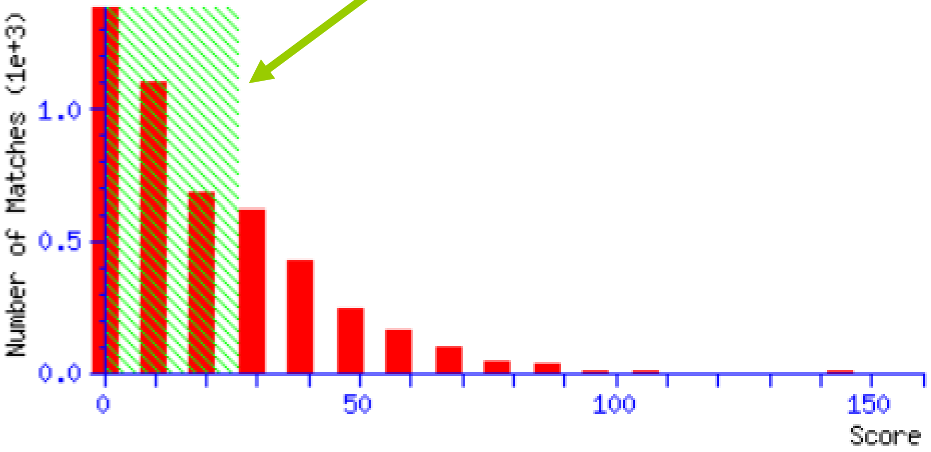
Mascot scoring

Mascot Score Histogram

Peptide score distribution.

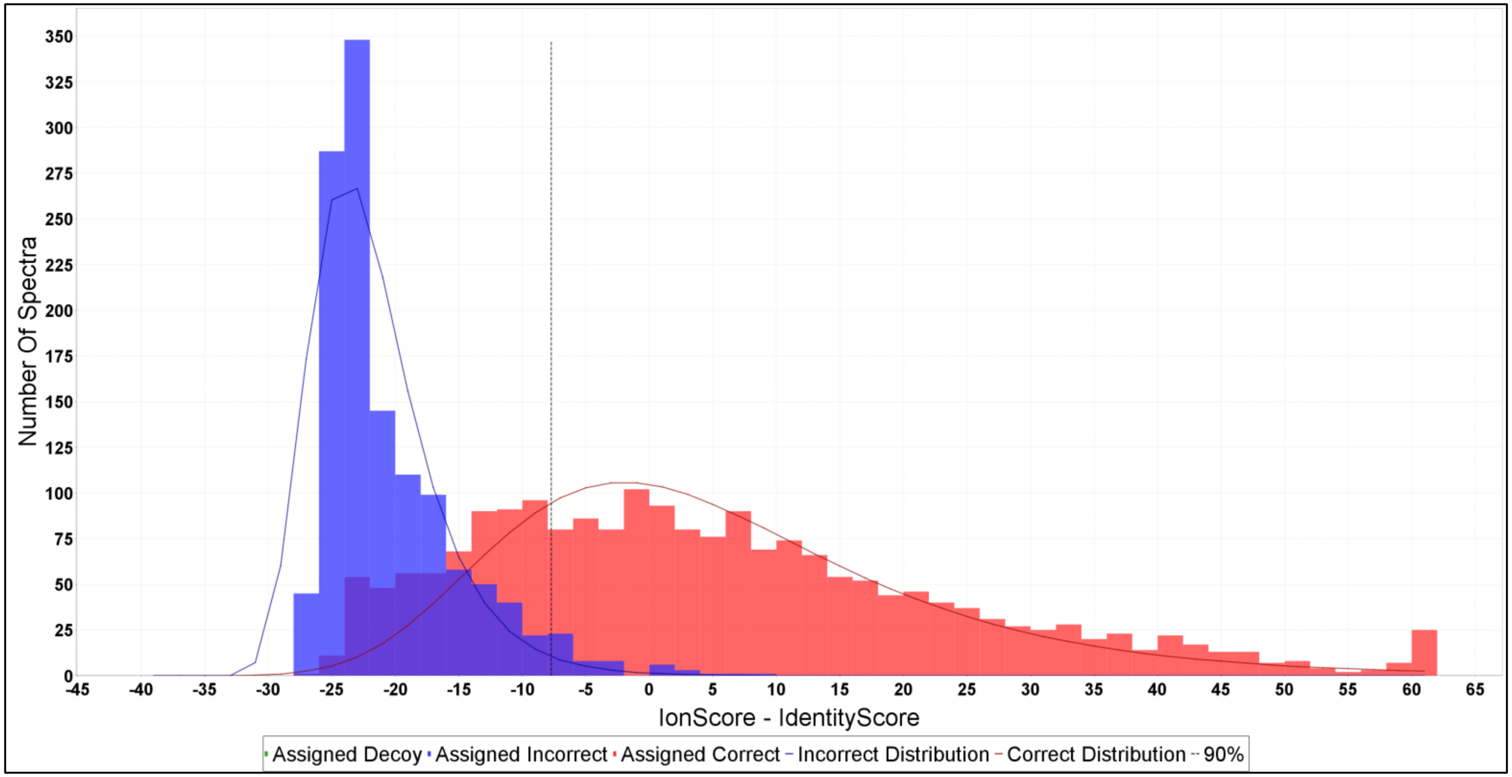
Ions score is −10log(*P*), where *P* is the probability that the observed match is a random event.

On average, individual ions scores **> 26** (beyond green shading) indicate **identity or extensive homology** (p<0.05).

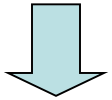


- At threshold score there is a 5% probability of random peptide spectrum match (PSM)
- When identifying several thousands PSMs, a significant number of them are random matches: multiple testing problem
- How many wrong identifications, False Discovery Rate (FDR) ?

Spectra score distribution

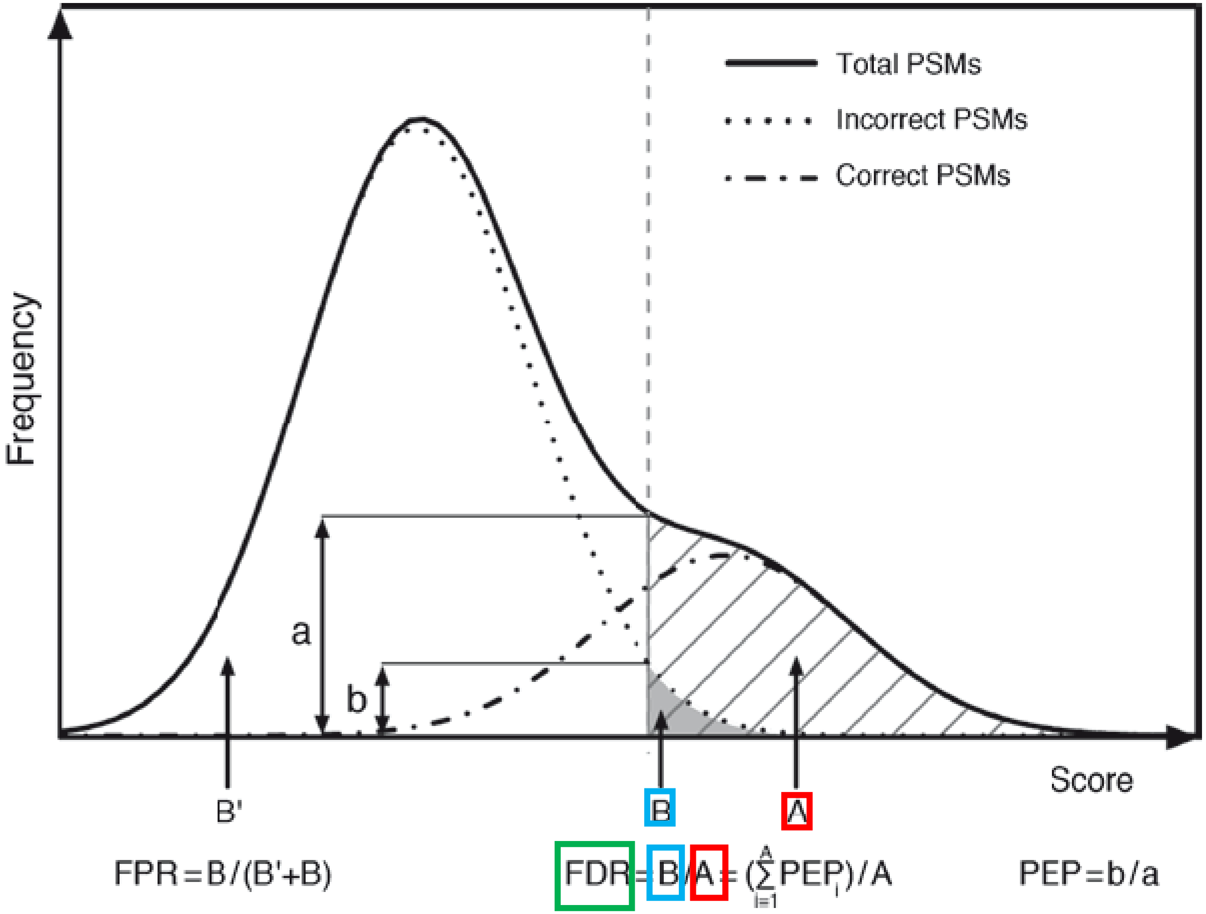


- Score distribution can be modeled for spectra validation (ex: **Peptide Prophet**)



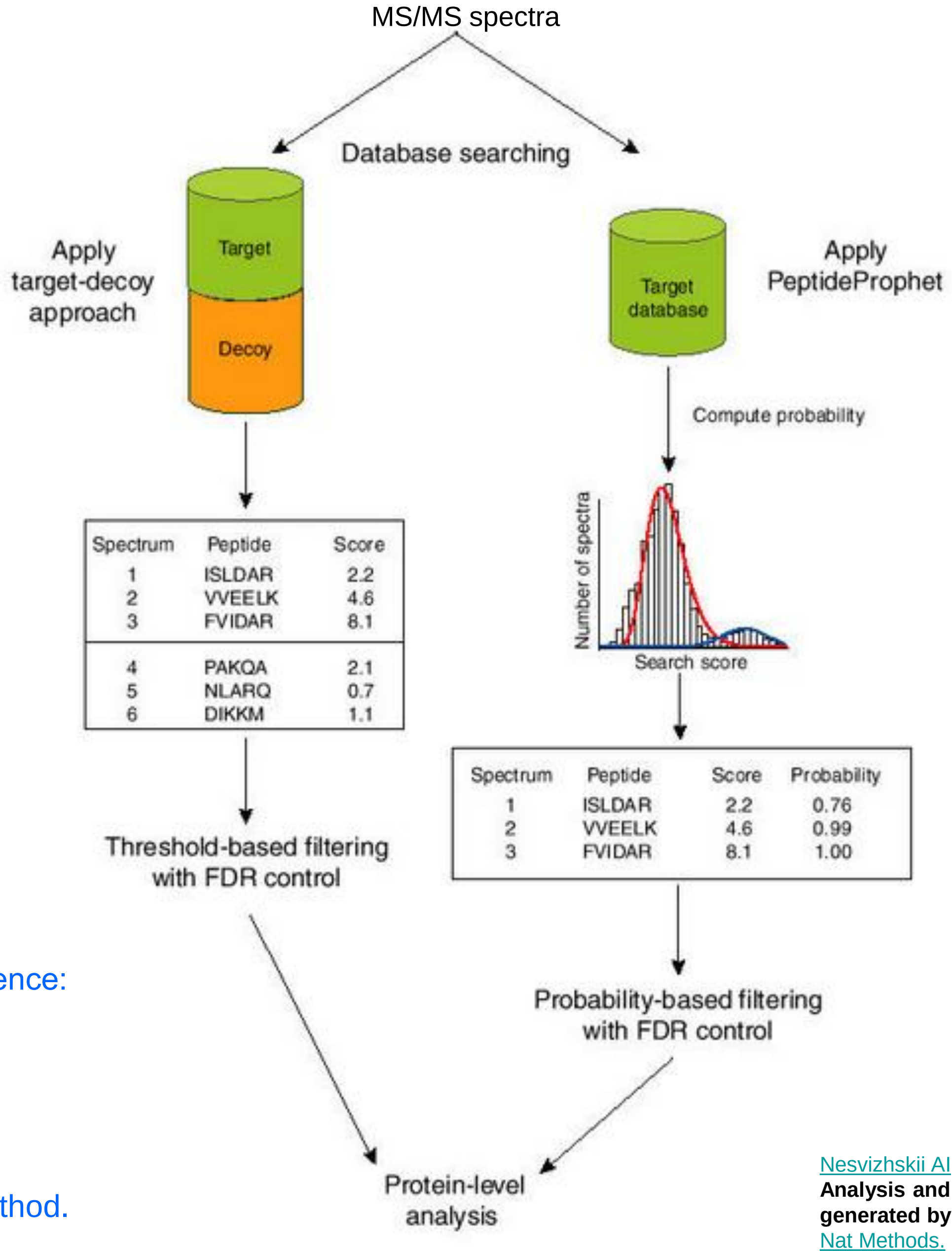
FDR

False Discovery Rate (FDR) - Peptides



FDR is defined as the expected proportion of incorrect identification among all identifications judged correct

Target-Decoy and probability-based filtering



MS/MS spectra

Database searching

Apply target-decoy approach

Spectrum	Peptide	Score
1	ISLDAR	2.2
2	VVEELK	4.6
3	FVIDAR	8.1
4	PAKQA	2.1
5	NLARQ	0.7
6	DIKKM	1.1

Threshold-based filtering with FDR control

Apply PeptideProphet

Compute probability

Spectrum	Peptide	Score	Probability
1	ISLDAR	2.2	0.76
2	VVEELK	4.6	0.99
3	FVIDAR	8.1	1.00

Probability-based filtering with FDR control

Protein-level analysis

→ Decoy DB is built by sequence:

- reversing
- shuffling
- randomization

or a combination of either method.

Nesvizhskii AI[1], Vitek O, Aebersold R.
Analysis and validation of proteomic data generated by tandem mass spectrometry.
Nat Methods. 2007 Oct;4(10):787-97.

False Discovery Rate - summary

- FDR is calculated at the peptide and at the protein (ex: Protein Prophet) level
- Various approach exist for FDR calculation, most of them relying on target-decoy approach
- Reporting of the FDR threshold applied is mandatory in proteomics publication: usually 1% is selected at peptide and protein level
- For very large datasets, FDR calculation is challenging and specific algorithms must be applied

Caveat

Protein identification

IS NOT

protein characterisation

Two peptides are enough to identify a protein
but
we are still identifying two peptides, not the entire protein

Highly similar sequences cannot be distinguished

For finding PTMs extensive sequence coverage is essential !!!

Post-translational modifications

Modification proteomics: the analysis of post-translational modifications (PTMs).

Typical question : how is protein activity modulated by covalent chemical modification ?

Some common PTMs

Modification	Δ Mass	Residue	Origin
Proteolysis	Various	Any	PTM, artefact
Dehydration	- 18.0106	N, Q, S, T, Y	PTM, artefact
Glycosylation (N-, O-, simple/complex)	Various	N, S, T, (Q)	PTM
Phosphorylation	+ 79.9663	S, T, Y	PTM
Sulfonation	+ 79.9568	S,T,Y,C	PTM
Acetylation	+ 42.0106	N-term or K	PTM, derivative
Carbamidomethylation	+ 57.0215	C	Derivative
Methylation	+ 14.0156	K, R, D, E, …	PTM, artefact
Ubiquitination (mono-, di-, poly, K48, K63, ..)	Various / + 114.043	K	PTM
Sumoylation (SUMO-1, -2, -3)	Various	K	PTM
Oxidation	+ 15.9949	C, M, W	PTM, artefact
ADP-ribosylation	+ 541.0611	R,C,N,S,E	PTM
Myristoylation	+ 210.1984	N-term G, K, C	PTM
Palmitoylation	+ 238.2297	C, K, S, T, N-term	PTM
Prenylation (farnesyl-, geranylgeranyl-)	Various	CaaX (C-term)	PTM
Nitrosylation	+ 28.9902	C	PTM
….. Almost 100 known….			

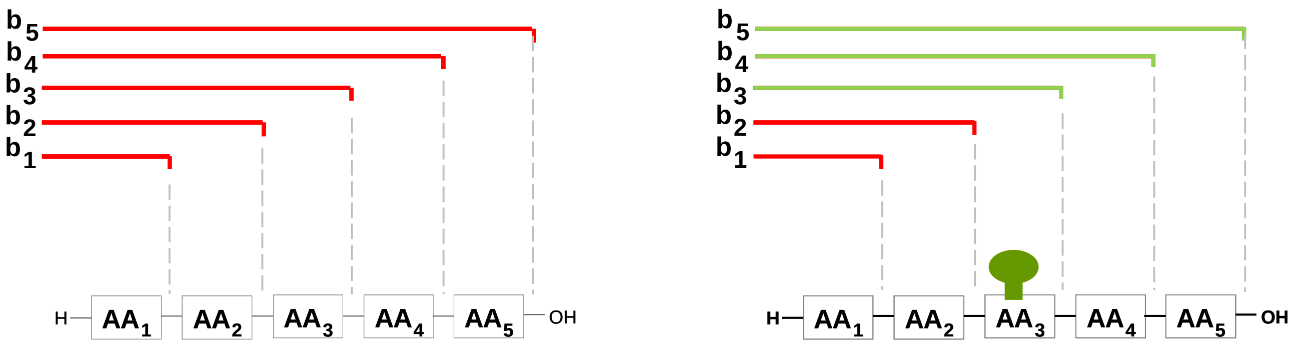
Mass shifts induced by Post Translational Modifications (PTMs)

Native peptide

M

Modified peptide

M*=M+ΔM



Fragment masses

b3*=b3+ΔM

b4*=b4+ΔM

b5*=b5+ΔM

MS analysis and modified matching parameters can identify modified peptides and sites of modification

Common issues in PTM analysis

- Protein sequence coverage
  → can be increased by multi-enzyme digestion, linked to abundance issue

- Labile PTMs / MS suitability
  → enzyme inhibitors, PTM derivatisation, use of alternative MS fragmentation (for ex. ETD)

- Abundance
  → PTM enrichment

- Artefacts
  → appropriate sample preparation, control experiments

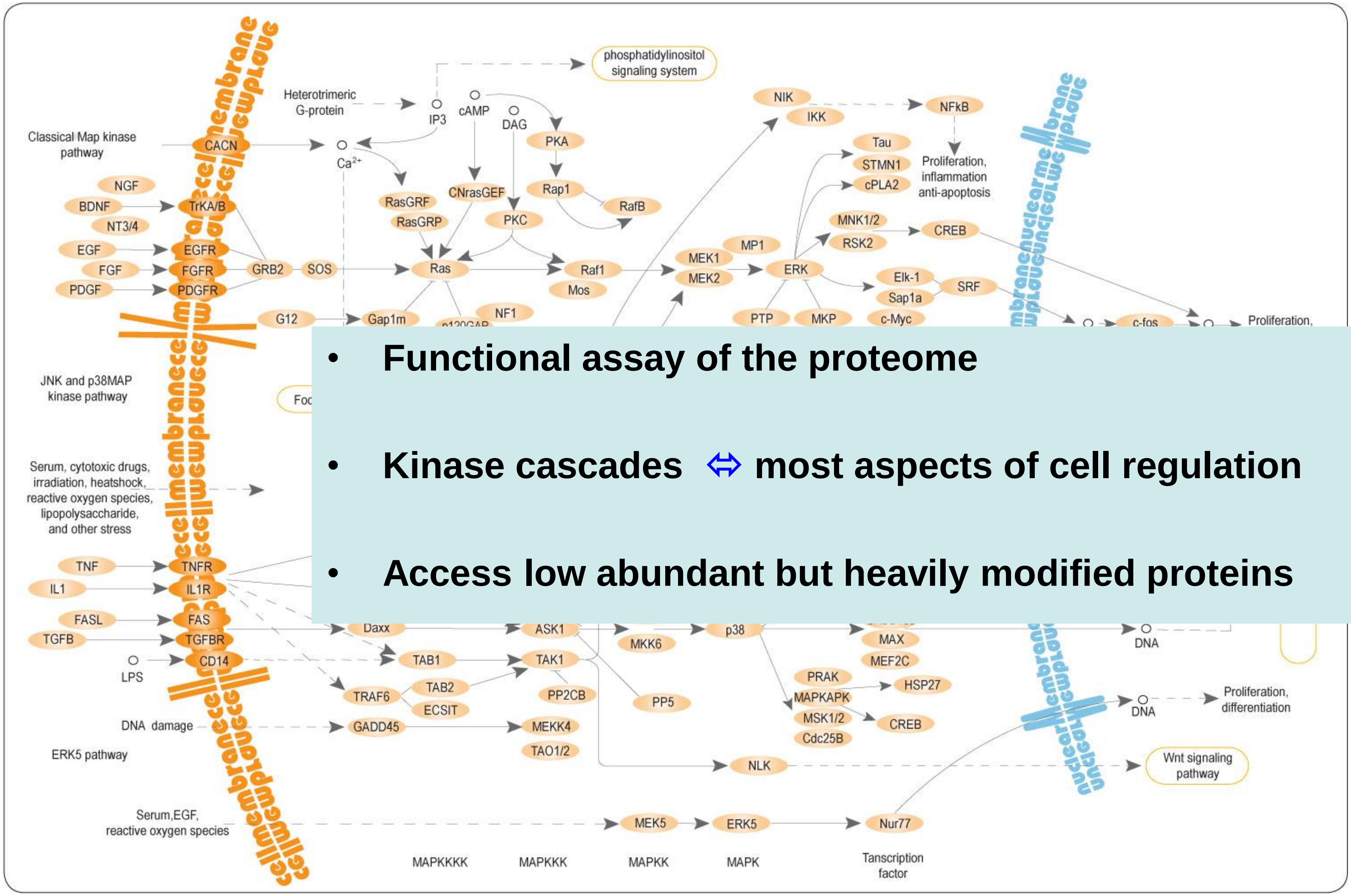
- Isobaric PTMs
  → high resolution MS, specific fragments (for ex. immonium ions)

- Unknown (untargeted) PTMs
  → error-tolerant search, blind search

- Localization
  → use of alternative MS fragmentation, localization algorithms

- Connectivity
  → middle-down / top-down analysis

Phosphoproteomics



- **Functional assay of the proteome**
- **Kinase cascades ⇔ most aspects of cell regulation**
- **Access low abundant but heavily modified proteins**

From : Wikimedia commons

Questions in phosphorylation analysis

- Is a protein of interest phosphorylated ?
- Which proteins are phosphorylated in a cell
  (or in a precise pathway ?)
- Localizing phosphorylation sites : exact residues
- Quantitation of changes in response to a stimulus
- Effect on physiological protein activity

Problems with phosphopeptide analysis

1) Quantity problem : abundance of the protein to analyze is often low and phosphorylation is substoichiometric, especially when purifying from in vivo

→ Scale up preparation, P-peptide enrichment

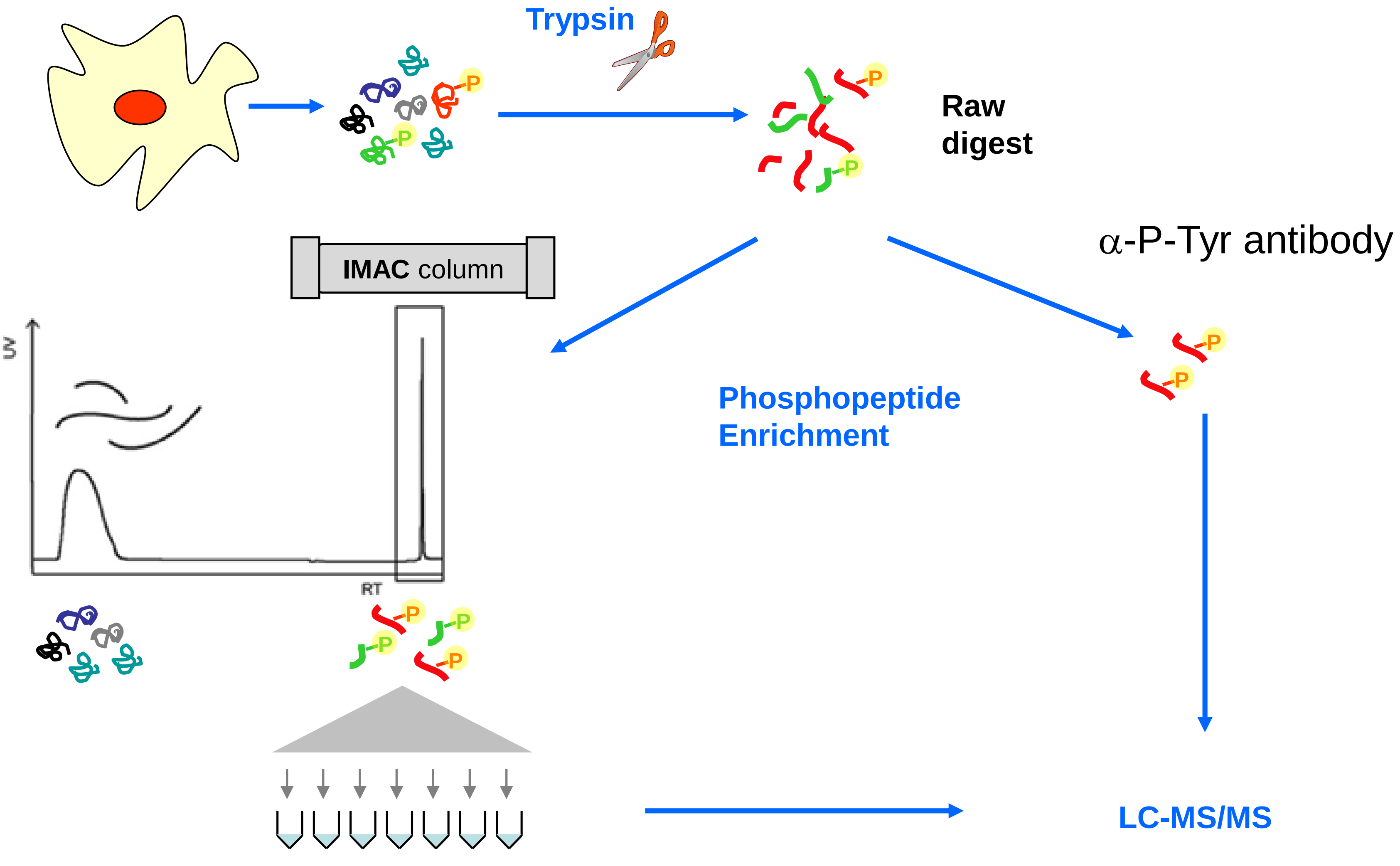
2) Bad fragmentation due to neutral loss : highly variable depending on peptide sequence

→ Choice of MS instrument and MS/MS fragmentation (HCD, CID, ETD)

3) Enzyme used for digestion: is trypsin always the correct choice ?
Phosphorylated regions are sometimes (often ?) in problematic regions of proteins : very acidic, K/R-poor or K/R-rich sequences

→ Digestion with multiple proteases

Typical workflows for phosphopeptide enrichment



Trypsin

Raw digest

α-P-Tyr antibody

IMAC column

Phosphopeptide Enrichment

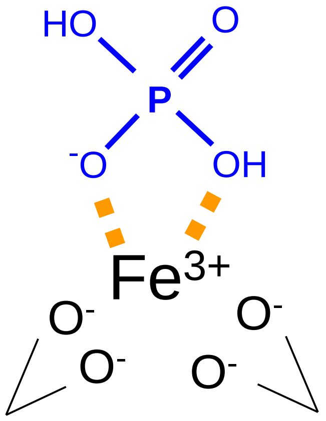
LC-MS/MS

P-Peptide fractionation (SCX, basic RP, …)

ID : P-proteins
P-peptides
P-sites

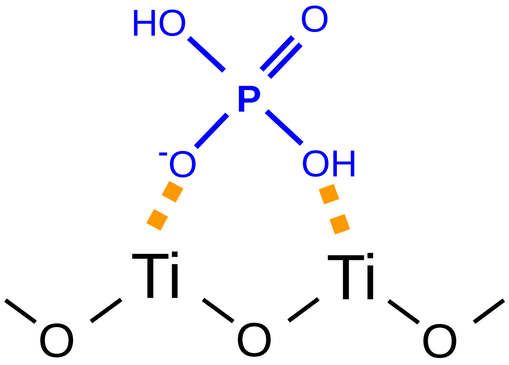
Enrichment of P-peptides by affinity chromatography

Classical IMAC :
Chelated Fe3+
(ex. IDA : iminodiacetic acid)



- Bind : pH 3-4
  Aqueous conditions
- Wash : pH 3-4
  Aqueous conditions
- Release : pH >9

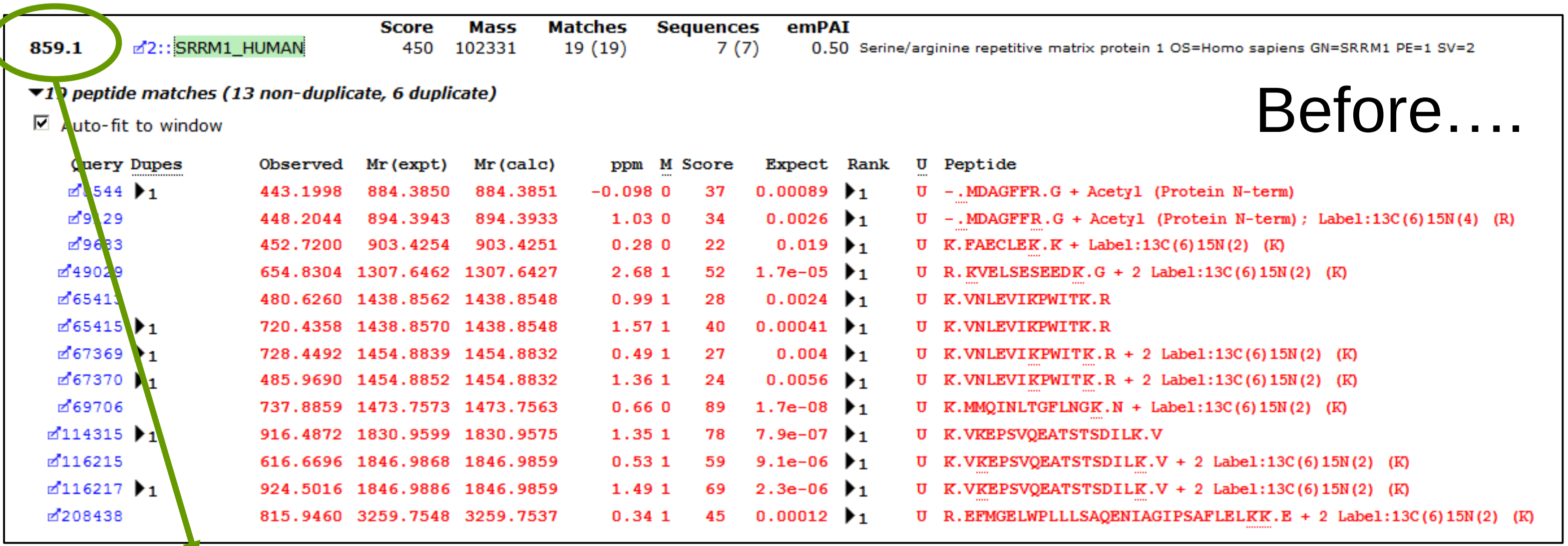
Metal oxides :
Titanium dioxide
Aluminium hydroxide
Zirconium oxides



- Bind : pH 1.5-4
  Almost any solvent
- Wash : pH 1.5-4
  Almost any solvent
- Release : pH >9

Exact binding mode ?

Snapshot : P-peptides enrichment by IMAC



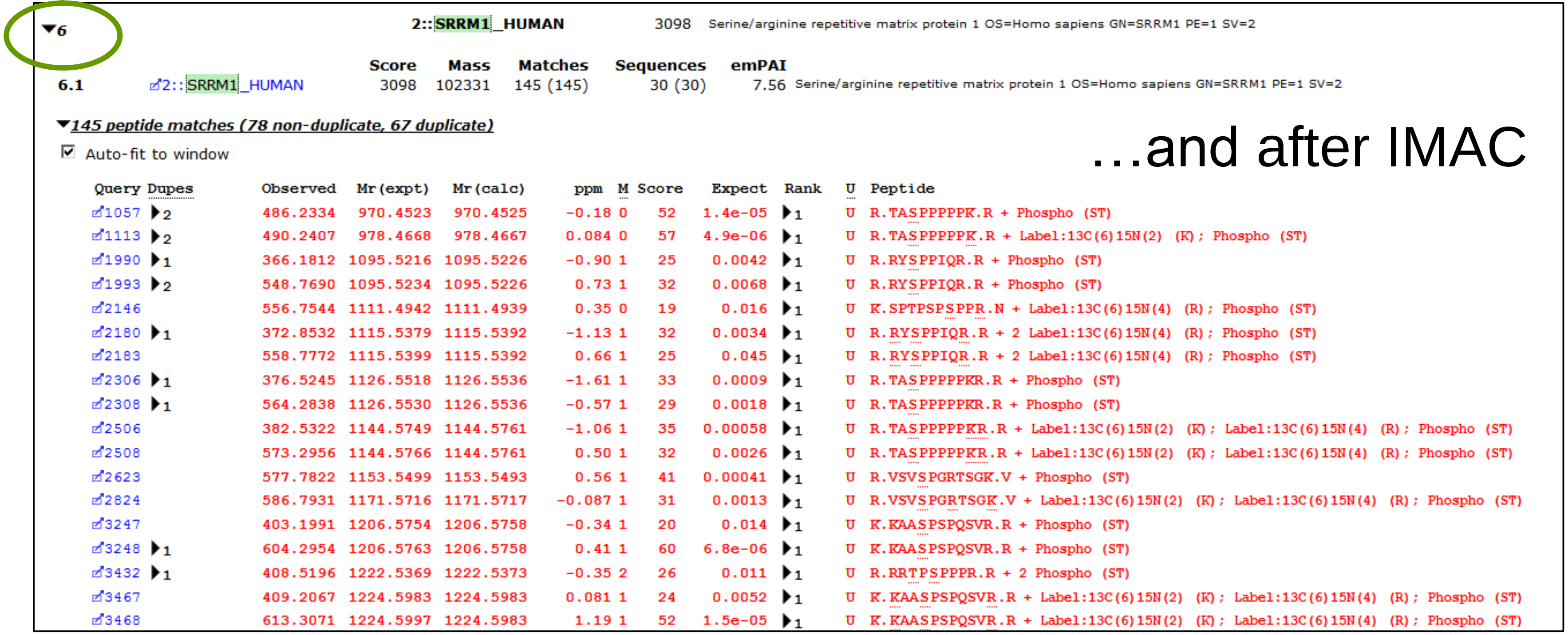
Before….

859.1	2::SRRM1_HUMAN	Score	Mass	Matches	Sequences	emPAI	
		450	102331	19 (19)	7 (7)	0.50	Serine/arginine repetitive matrix protein 1 OS=Homo sapiens GN=SRRM1 PE=1 SV=2

▼19 peptide matches (13 non-duplicate, 6 duplicate)

☑ Auto-fit to window

Query	Dupes	Observed	Mr(expt)	Mr(calc)	ppm	M	Score	Expect	Rank	U	Peptide
8544	▶1	443.1998	884.3850	884.3851	-0.098	0	37	0.00089	▶1	U	-.MDAGFFR.G + Acetyl (Protein N-term)
9129		448.2044	894.3943	894.3933	1.03	0	34	0.0026	▶1	U	-.MDAGFFR.G + Acetyl (Protein N-term); Label:13C(6)15N(4) (R)
9683		452.7200	903.4254	903.4251	0.28	0	22	0.019	▶1	U	K.FAECLEK.K + Label:13C(6)15N(2) (K)
49029		654.8304	1307.6462	1307.6427	2.68	1	52	1.7e-05	▶1	U	R.KVELSESEEDK.G + 2 Label:13C(6)15N(2) (K)
65413		480.6260	1438.8562	1438.8548	0.99	1	28	0.0024	▶1	U	K.VNLEVIKPWITK.R
65415	▶1	720.4358	1438.8570	1438.8548	1.57	1	40	0.00041	▶1	U	K.VNLEVIKPWITK.R
67369	▶1	728.4492	1454.8839	1454.8832	0.49	1	27	0.004	▶1	U	K.VNLEVIKPWITK.R + 2 Label:13C(6)15N(2) (K)
67370	▶1	485.9690	1454.8852	1454.8832	1.36	1	24	0.0056	▶1	U	K.VNLEVIKPWITK.R + 2 Label:13C(6)15N(2) (K)
69706		737.8859	1473.7573	1473.7563	0.66	0	89	1.7e-08	▶1	U	K.MMQINLTGFLNGK.N + Label:13C(6)15N(2) (K)
114315	▶1	916.4872	1830.9599	1830.9575	1.35	1	78	7.9e-07	▶1	U	K.VKEPSVQEATSTSDILK.V
116215		616.6696	1846.9868	1846.9859	0.53	1	59	9.1e-06	▶1	U	K.VKEPSVQEATSTSDILK.V + 2 Label:13C(6)15N(2) (K)
116217	▶1	924.5016	1846.9886	1846.9859	1.49	1	69	2.3e-06	▶1	U	K.VKEPSVQEATSTSDILK.V + 2 Label:13C(6)15N(2) (K)
208438		815.9460	3259.7548	3259.7537	0.34	1	45	0.00012	▶1	U	R.EFMGELWPLLLSAQENIAGIPSAFLELKK.E + 2 Label:13C(6)15N(2) (K)



…and after IMAC

▼6 2::SRRM1_HUMAN 3098 Serine/arginine repetitive matrix protein 1 OS=Homo sapiens GN=SRRM1 PE=1 SV=2

6.1	2::SRRM1_HUMAN	Score	Mass	Matches	Sequences	emPAI	
		3098	102331	145 (145)	30 (30)	7.56	Serine/arginine repetitive matrix protein 1 OS=Homo sapiens GN=SRRM1 PE=1 SV=2

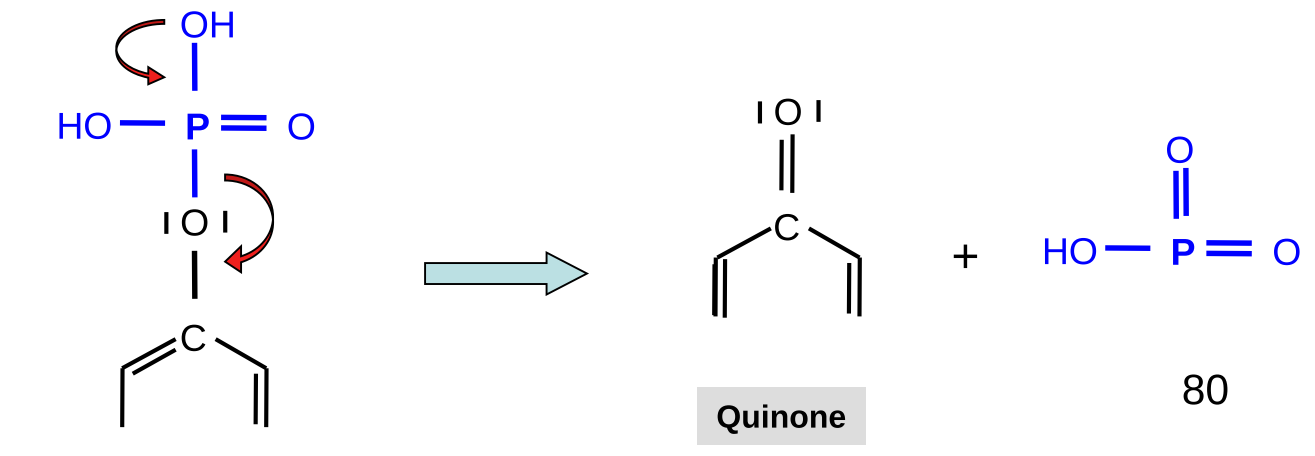
▼145 peptide matches (78 non-duplicate, 67 duplicate)

☑ Auto-fit to window

Query	Dupes	Observed	Mr(expt)	Mr(calc)	ppm	M	Score	Expect	Rank	U	Peptide
1057	▶2	486.2334	970.4523	970.4525	-0.18	0	52	1.4e-05	▶1	U	R.TASPPPPPK.R + Phospho (ST)
1113	▶2	490.2407	978.4668	978.4667	0.084	0	57	4.9e-06	▶1	U	R.TASPPPPPK.R + Label:13C(6)15N(2) (K); Phospho (ST)
1990	▶1	366.1812	1095.5216	1095.5226	-0.90	1	25	0.0042	▶1	U	R.RYSPPIQR.R + Phospho (ST)
1993	▶2	548.7690	1095.5234	1095.5226	0.73	1	32	0.0068	▶1	U	R.RYSPPIQR.R + Phospho (ST)
2146		556.7544	1111.4942	1111.4939	0.35	0	19	0.016	▶1	U	K.SPTPSPSPPR.N + Label:13C(6)15N(4) (R); Phospho (ST)
2180	▶1	372.8532	1115.5379	1115.5392	-1.13	1	32	0.0034	▶1	U	R.RYSPPIQR.R + 2 Label:13C(6)15N(4) (R); Phospho (ST)
2183		558.7772	1115.5399	1115.5392	0.66	1	25	0.045	▶1	U	R.RYSPPIQR.R + 2 Label:13C(6)15N(4) (R); Phospho (ST)
2306	▶1	376.5245	1126.5518	1126.5536	-1.61	1	33	0.0009	▶1	U	R.TASPPPPPKR.R + Phospho (ST)
2308	▶1	564.2838	1126.5530	1126.5536	-0.57	1	29	0.0018	▶1	U	R.TASPPPPPKR.R + Phospho (ST)
2506		382.5322	1144.5749	1144.5761	-1.06	1	35	0.00058	▶1	U	R.TASPPPPPKR.R + Label:13C(6)15N(2) (K); Label:13C(6)15N(4) (R); Phospho (ST)
2508		573.2956	1144.5766	1144.5761	0.50	1	32	0.0026	▶1	U	R.TASPPPPPKR.R + Label:13C(6)15N(2) (K); Label:13C(6)15N(4) (R); Phospho (ST)
2623		577.7822	1153.5499	1153.5493	0.56	1	41	0.00041	▶1	U	R.VSVSPGRTSGK.V + Phospho (ST)
2824		586.7931	1171.5716	1171.5717	-0.087	1	31	0.0013	▶1	U	R.VSVSPGRTSGK.V + Label:13C(6)15N(2) (K); Label:13C(6)15N(4) (R); Phospho (ST)
3247		403.1991	1206.5754	1206.5758	-0.34	1	20	0.014	▶1	U	K.KAASPSPQSVR.R + Phospho (ST)
3248	▶1	604.2954	1206.5763	1206.5758	0.41	1	60	6.8e-06	▶1	U	K.KAASPSPQSVR.R + Phospho (ST)
3432	▶1	408.5196	1222.5369	1222.5373	-0.35	2	26	0.011	▶1	U	R.RRTPSPPPR.R + 2 Phospho (ST)
3467		409.2067	1224.5983	1224.5983	0.081	1	24	0.0052	▶1	U	K.KAASPSPQSVR.R + Label:13C(6)15N(2) (K); Label:13C(6)15N(4) (R); Phospho (ST)
3468		613.3071	1224.5997	1224.5983	1.19	1	52	1.5e-05	▶1	U	K.KAASPSPQSVR.R + Label:13C(6)15N(2) (K); Label:13C(6)15N(4) (R); Phospho (ST)

Fragmentation (CID/HCD) reactions in positive mode

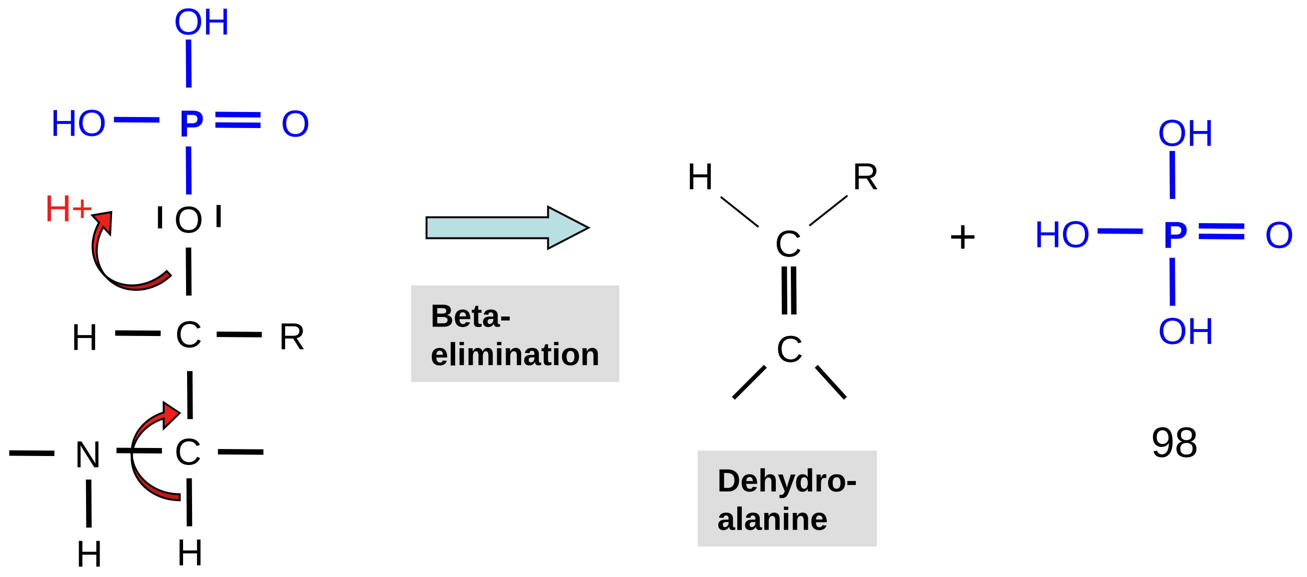
Tyr



Quinone

80

Ser
Thr



Beta-elimination

Dehydro-alanine

98

Multi-enzymatic strategy

- Protein sequence coverage can be improved by using different digestion enzymes:
  - ➢ Trypsin (K,R)
  - ➢ Chymotrypsin (F, L, W, Y)
  - ➢ Lys-C (K)
  - ➢ Gluc-C (D, E)
  - ➢ Arg-C (R)
  - ➢ Combination of 2 enzymes

Ex: POM1_SCHPO (*S. pombe*, fission yeast)

```
   1 MGYLQSQKAV SLGDENTDAL FKLHTSNRKS ANMFGIKSEL LNPSELSAVG
  51 SYSNDICPNR QSSSSTAADT SPSTNASNTN ISFPEQEHKD ELFMNVEPKG
 101 VGSSMDNHAI TIHHSTGNGL LRSSFDHDYR QKNSPRNSIH RLSNISIGNN
 151 PIDFESSQQN NPSSLNTSSH HRTSSISNSK SFGTSLSYYN RSSKPSDWNQ
 201 QNNGGHLSGV ISITQDVSSV PLQSSVFSSG NHAYHASMAP KRSGSWRHTN
 251 FHSTSHPRAA SIGNKSGIPP VPTIPPNIGH STDHQHPKAN ISGSLTKSSS
 301 ESKNLSTIQS PLKTSNSFFK ELSPHSQITL SNVKNNHSHV GSQTKSHSFA
 351 TPSVFDNNKP VSSDNHNNTT TSSQVHPDSR NPDPKAAPKA VSQKTNVDGH
 401 RNHEAKHGNT VQNESKSQKS SNKEGRSSRG GFFSRLSFSR SSSRMKKGSK
 451 AKHEDAPDVP AIPHAYIADS STKSSYRNGK KTPTRTKSRM QQFINWFKPS
 501 KERSSNGNSD SASPPPVPRL SITRSQVSRE PEKPEEIPSV PPLPSNFKDK
 551 GHVPQQRSVS YTPKRSSDTS ESLQPSLSFA SSNVLSEPFD RKVADLAMKA
 601 INSKRINKLL DDAKVMQSLL DRACIITPVR NTEVQLINTA PLTEYEQDEI
 651 NNYDNIYFTG LRNVDKRRSA DENTSSNFGF DDERGDYKVV LGDHIAYRYE
 701 VVDFLGKGSF GQVLRCIDYE TGKLVALKII RNKKRFHMQA LVETKILQKI
 751 REWDPLDEYC MVQYTDHFYF RDHLCVATEL LGKNLYELIK SNGFKGLPIV
 801 VIKSITRQLI QCLTLLNEKH VIHCDLKPEN ILLCHPFKSQ VKVIDFGSSC
 851 FEGECVYTYI QSRFYRSPEV ILGMGYGTPI DVWSLGCIIA EMYTGFPLFP
 901 GENEQEQLAC IMEIFGPPDH SLIDKCSRKK VFFDSSGKPR PFVSSKGVSR
 951 RPFSKSLHQV LQCKDVSFLS FISDCLKWDP DERMTPQQAA QHDFLTGKQD
1001 VRRPNTAPAR QKFARPPNIE TAPIPRPLPN LPMEYNDHTL PSPKEPSNQA
1051 SNLVRSSDKF PNLITNLDYS IISDNGFLRK PVEKSRP
```

Semi-specific search, 4 missed cleavages allowed:

1. SEQ: sequence covered with trypsin digestion

2. SEQ: additional sequence covered with chymotrypsin digestion

3. SEQ: additional sequence covered with Lys-C digestion

4. SEQ: additional sequence covered with Glu-C digestion

=> Total sequence coverage: 95.9 %

S: phosphosite found with trypsin

S: additional phosphosite found with chymotrypsin

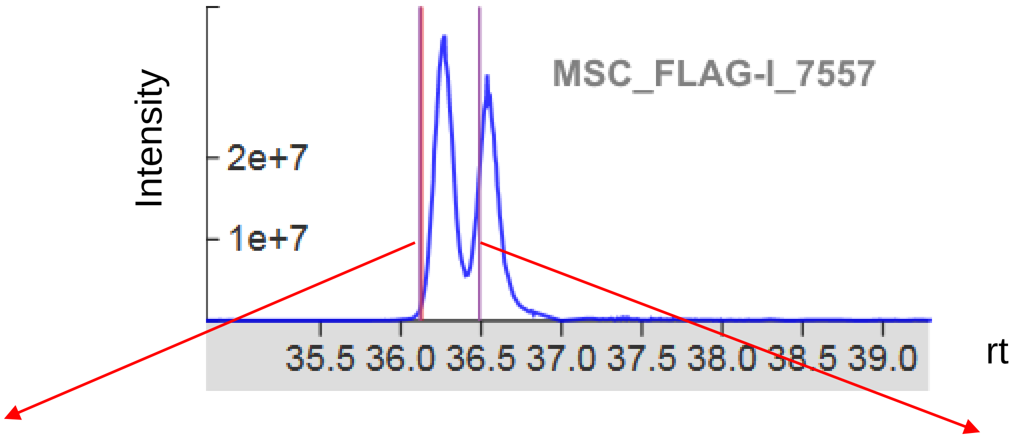
S: additional phosphosite found with Lys-C

SS: ambiguous phosphosite localization

=> Total number of phosphosites : 41

Phosphorylation localization

Extracted Ion Chromatogram (XIC) of phosphorylated DLHQPSLSPASPHSQGFER (*m/z* 723.996)



DLHQPSLS{Phospho}PASPHSQGFER

DLHQPSLSPAS{Phospho}PHSQGFER

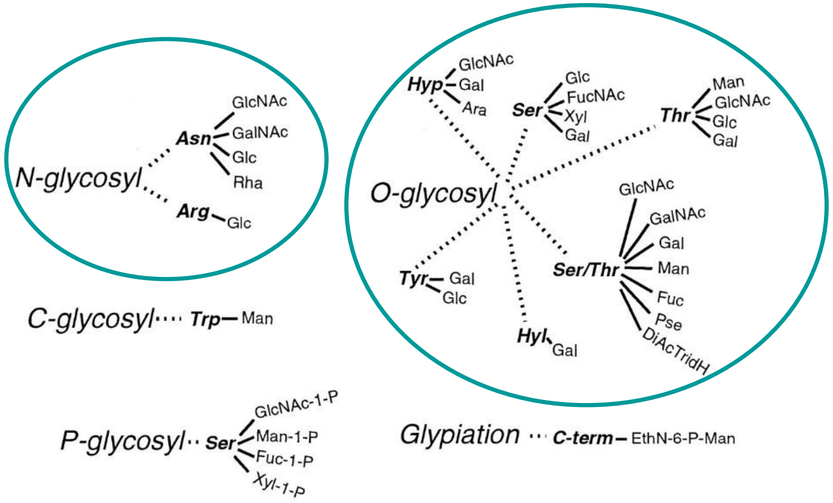
MD-score 85.8% 86.2%

- Various phosphoforms of the same P-peptide can be sometimes distinguished by their different retention times
- Localization algorithms (Ascore, ptmRS, LuciPHOr, Mascot Delta Score, …) for automated site assignment with probability score
- In case of phosphoform co-elution, site discrimination (and quantification) is often impossible !

PTM exercise

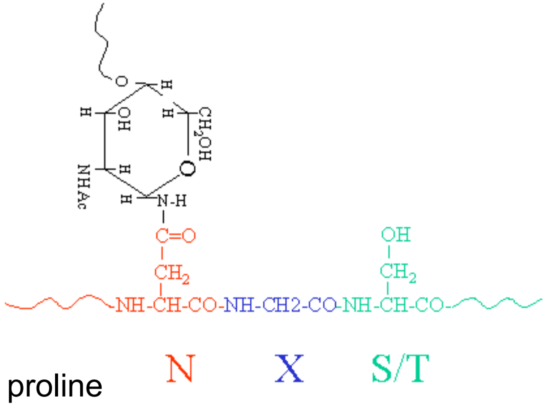
Glycosylation -1

Various glycosylation linkages:



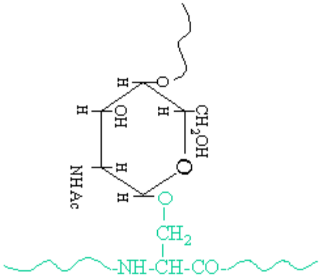
From: Spiro RG, Glycobiology 2002, 12:43R-56R

- **Consensus sequence for N-glycosylation:**



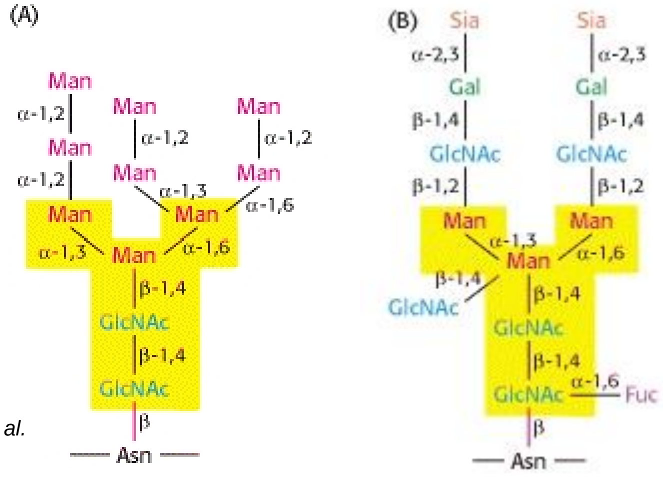
X ≠ proline

- **No consensus sequence for O-glycosylation:**



From: www.ionsource.com

Complexity & heterogeneity of glycans:

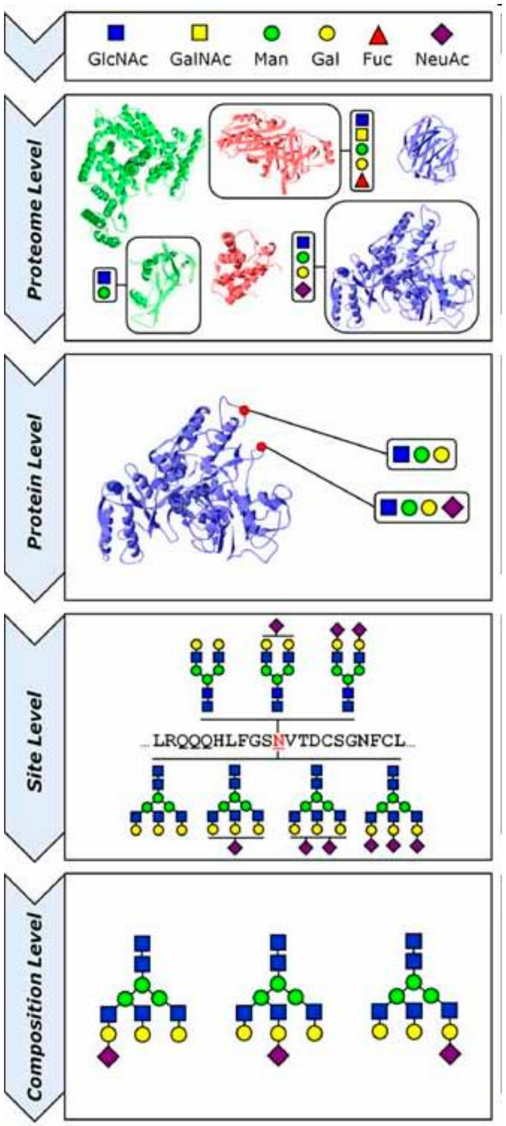


A pentasaccharide core (shaded yellow)
is common to all N-linked oligosaccharides
A) high-mannose type, (B) complex type

From: Biochemistry. 5th edition., Berg JM, et al.
New York: W H Freeman; 2002.

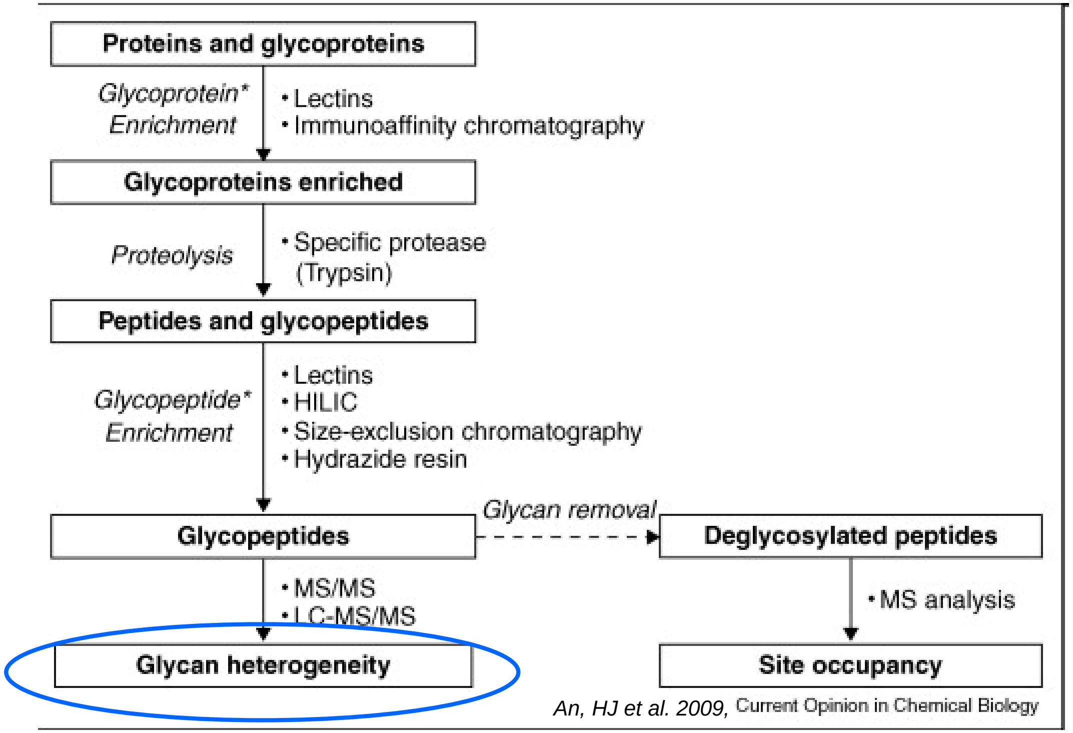
Glycosylation -2

Levels of glycoproteomic complexity



From: Dodds, 2012, Mass Spectrometry Reviews

Standard approaches to determine site-specific glycosylation



Proteins and glycoproteins

Glycoprotein Enrichment*
- Lectins
- Immunoaffinity chromatography

Glycoproteins enriched

Proteolysis
- Specific protease (Trypsin)

Peptides and glycopeptides

Glycopeptide Enrichment*
- Lectins
- HILIC
- Size-exclusion chromatography
- Hydrazide resin

Glycopeptides — *Glycan removal* → **Deglycosylated peptides**

- MS/MS
- LC-MS/MS

Glycan heterogeneity

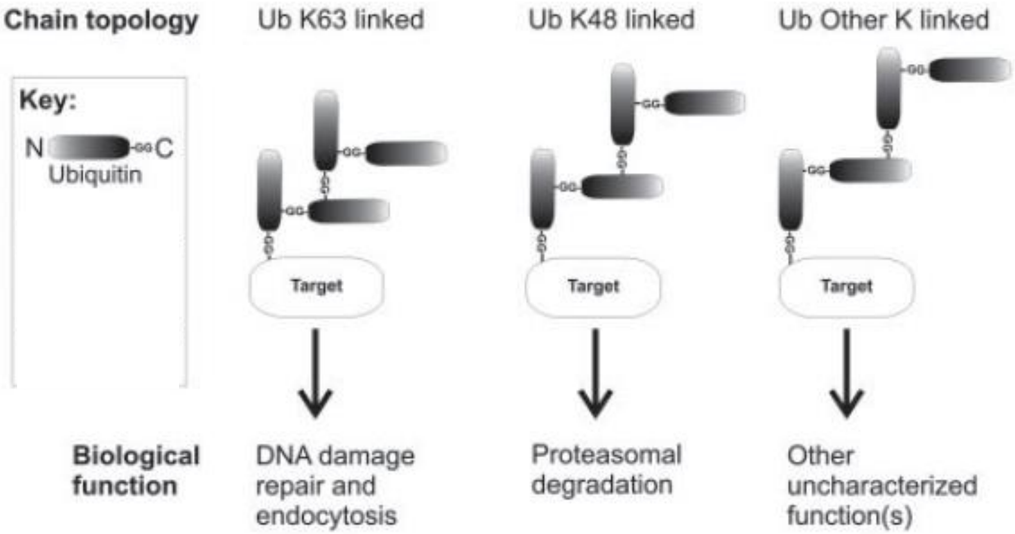
- MS analysis

Site occupancy

An, HJ et al. 2009, Current Opinion in Chemical Biology

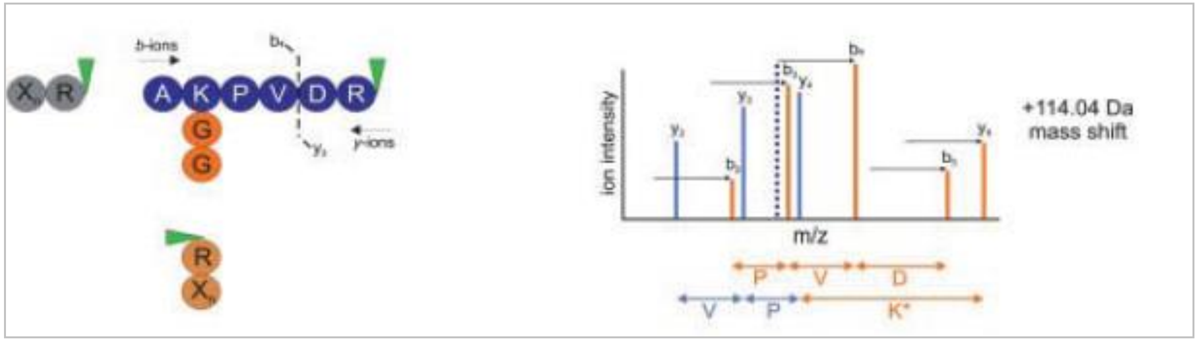
Glycomics

Ubiquitination



General questions :

- **Crosslinking site (target protein)**
- **Mono- or polyubiquitination ?**
- **Ub chain linkage type (48,63,…)**



For large scale studies :
Ub-modified proteins can be enriched by :

1) Expr. N-term tagged Ub => affinity purification

2) Antibodies or resins with immobilized Ub-binding domains (UIM,UBA,UBZ…)

After trypsin cleavage :
Peptide modification is +GlyGly, ΔM=+114.0429
Isobaric with +N, +2*Iodoacetamide

Using mass spectrometry to identify ubiquitin and ubiquitin-like protein conjugation sites
Stanley M. Jeram , Tharan Srikumar , Patrick G. A. Pedrioli and Brian Raught , Proteomics 2009, *9*, 922–934

Ub-like proteins

C-termini of some Human Ub-like proteins

Ubiquitin

…VLRLR**GG**

SUMO1

…TPKELGMEEEDVIEVYQEQT**GG**

SUMO2

…DTPAQLEMEDEDTIDVFQQQT**GG**

SUMO3

…DTPAQLEMEDEDTIDVFQQQT**GG**

NEDD8

…DYKILGGSVLHLVLALR**GG**

ISG15

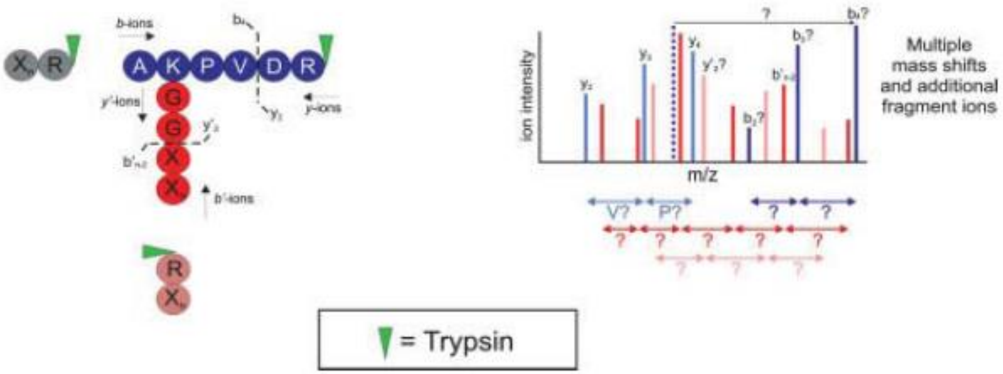
…GLKPLSTVFMNLRLR**GG**

URM1

….LGELDYQLQDQDSVLFISTLH**GG**

Problems :

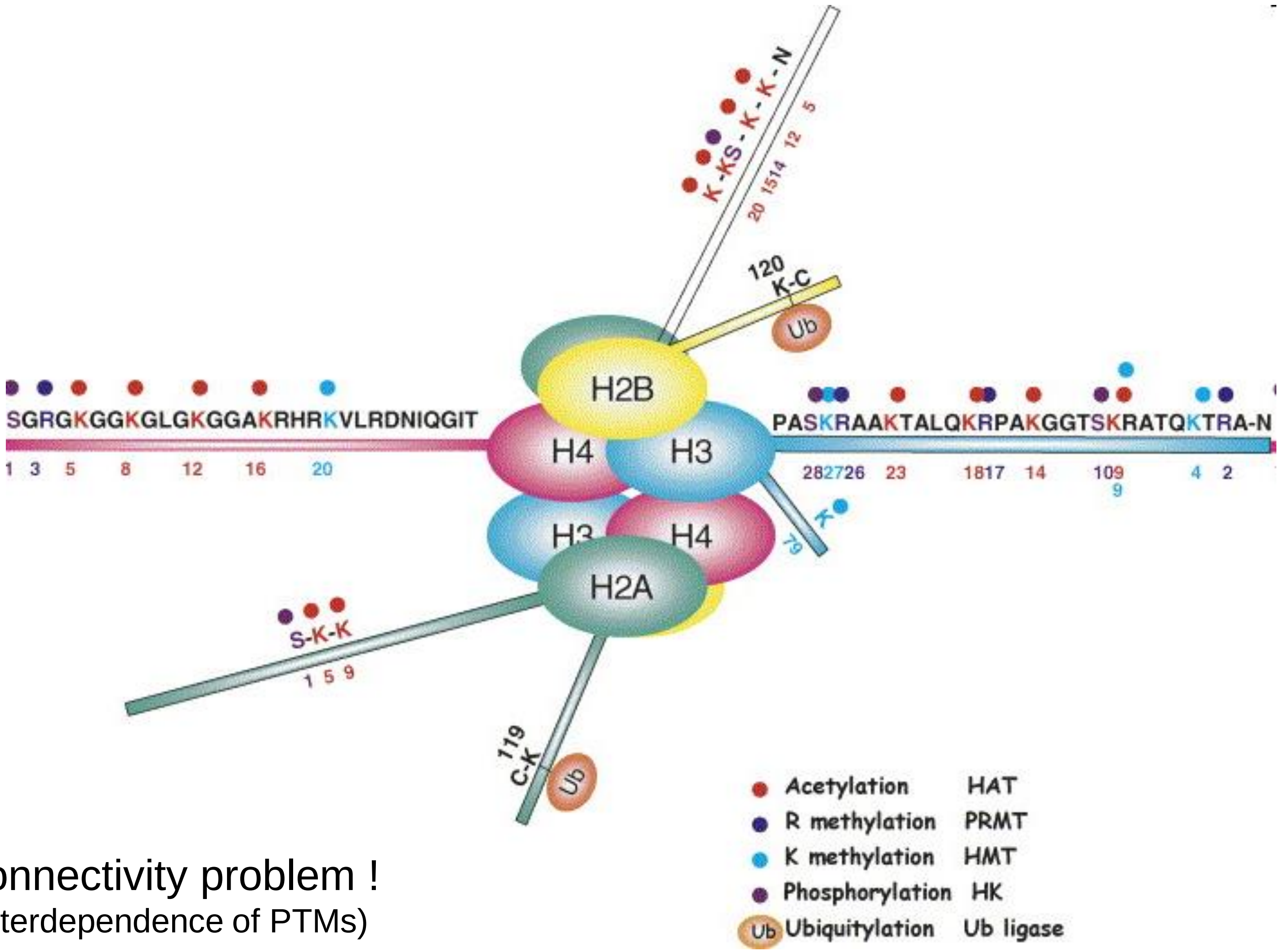
SUMO, URM1 :
Trypsin => Large cross linked peptides
Other proteases : poorer activity, specificity, MS of peptides
Fragmentation patterns complex, special software needed



NEW : WALP protease : ..T GG-

Using mass spectrometry to identify ubiquitin and ubiquitin-like protein conjugation sites
Stanley M. Jeram , Tharan Srikumar , Patrick G. A. Pedrioli and Brian Raught , Proteomics 2009, *9*, 922–934

The extreme case : histone tails



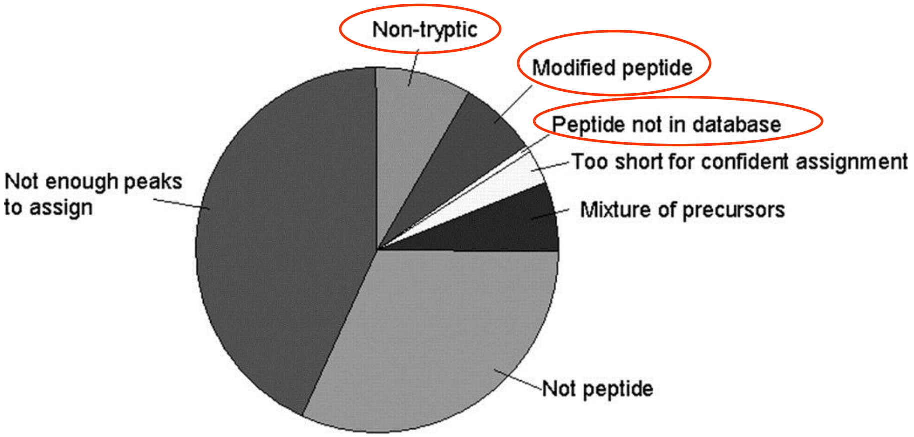
⇒ connectivity problem !
(interdependence of PTMs)

→ ***Middle-down proteomics***

Figure from: Altucci L. et al., 2005, Int. J. Biochem. Cell Biol., 37(9): 1752-62.

Unknown PTMs – mining unassigned MS/MS spectra

- A large proportion of MS/MS spectra are not assigned in proteomics samples:



Ex. From: Chalkley R J et al. Mol Cell Proteomics 2005;4:1189-1193

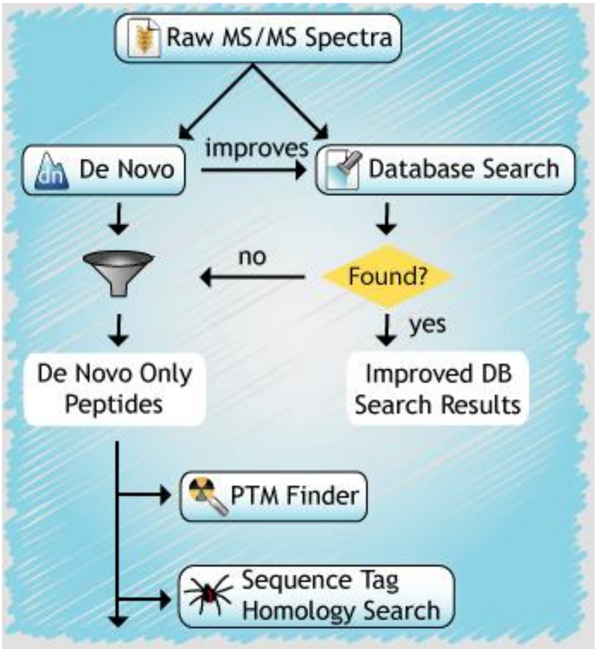
- Percentage of non-ID spectra is highly variable, 25-60%, depending on: sample complexity, MS instrument resolution, depth of analysis, DB search parameters, …

- All possible unknown PTMs cannot be searched in the classical way because of too large search space: → high % of false positive matches, computationally heavy

Mining Unknown PTMs

- Various strategies/software used for discovery of unknown PTMs: error-tolerant search (Mascot), dependent peptides (MaxQuant), MODa, SpecOMS, Open-pNovo, PTM Finder (PEAKS), …

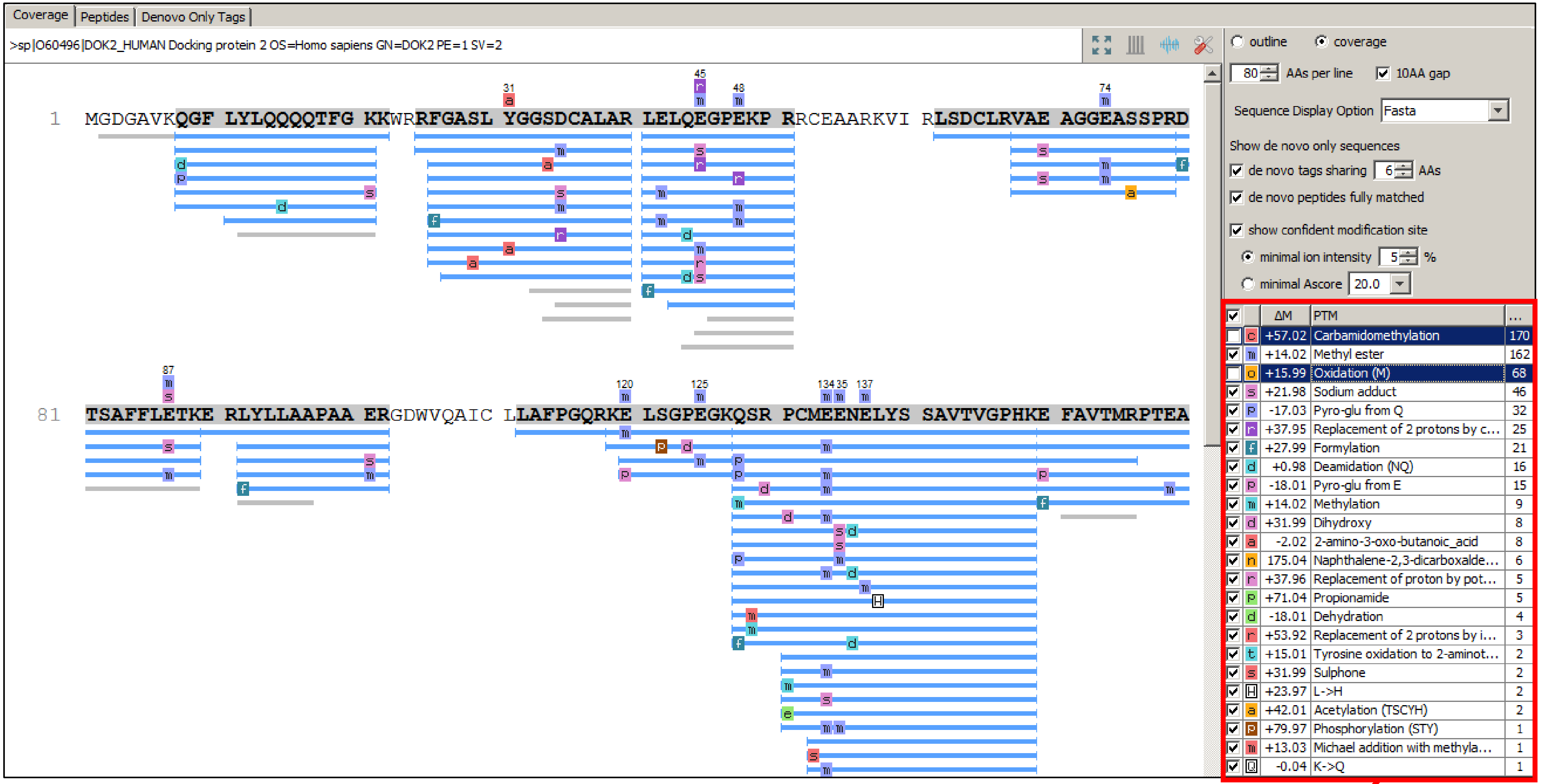
Ex: PEAKS workflow



- Results of open search must be interpreted with caution: many artefacts (PTM identity or position) !

- Many PTMs can be explained by sample preparation artefacts (oxidation, carbamylation, propionamide, …)

Ex: PEAKS PTM Finder



>sp|O60496|DOK2_HUMAN Docking protein 2 OS=Homo sapiens GN=DOK2 PE=1 SV=2

```
 1  MGDGAVKQGF LYLQQQQTFG KKWRRFGASL YGGSDCALAR LELQEGPEKP RRCEAARKVI RLSDCLRVAE AGGEASSPRD
81  TSAFFLETKE RLYLLAAPAA ERGDWVQAIC LLAFPGQRKE LSGPEGKQSR PCMEENELYS SAVTVGPHKE FAVTMRPTEA
```

○ outline ● coverage

80 AAs per line ☑ 10AA gap

Sequence Display Option: Fasta

Show de novo only sequences

☑ de novo tags sharing 6 AAs

☑ de novo peptides fully matched

☑ show confident modification site

● minimal ion intensity 5 %

○ minimal Ascore 20.0

☑		ΔM	PTM	...
□	c	+57.02	Carbamidomethylation	170
☑	m	+14.02	Methyl ester	162
□	o	+15.99	Oxidation (M)	68
☑	s	+21.98	Sodium adduct	46
☑	p	-17.03	Pyro-glu from Q	32
☑	r	+37.95	Replacement of 2 protons by c...	25
☑	f	+27.99	Formylation	21
☑	d	+0.98	Deamidation (NQ)	16
☑	p	-18.01	Pyro-glu from E	15
☑	m	+14.02	Methylation	9
☑	d	+31.99	Dihydroxy	8
☑	a	-2.02	2-amino-3-oxo-butanoic_acid	8
☑	n	175.04	Naphthalene-2,3-dicarboxalde...	6
☑	r	+37.96	Replacement of proton by pot...	5
☑	p	+71.04	Propionamide	5
☑	d	-18.01	Dehydration	4
☑	r	+53.92	Replacement of 2 protons by i...	3
☑	t	+15.01	Tyrosine oxidation to 2-aminot...	2
☑	s	+31.99	Sulphone	2
☑	H	+23.97	L->H	2
☑	a	+42.01	Acetylation (TSCYH)	2
☑	p	+79.97	Phosphorylation (STY)	1
☑	m	+13.03	Michael addition with methyla...	1
☑	Q	-0.04	K->Q	1

Additional PTMs found with open search

LUNCH

Quantitative proteomics

Expression proteomics : analysis of protein expression levels and their changes

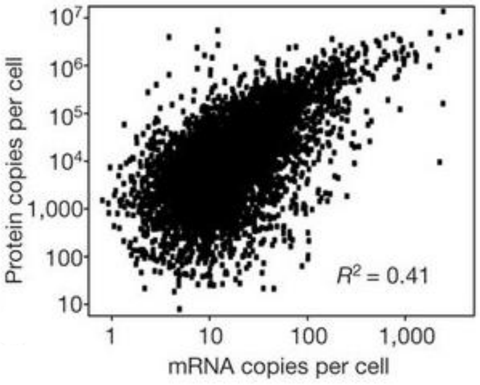
Typical questions :

**** what distinguishes a lymphocyte from a neuron ?***

**** which proteins are newly induced in a cell after a specific stimulus ?***

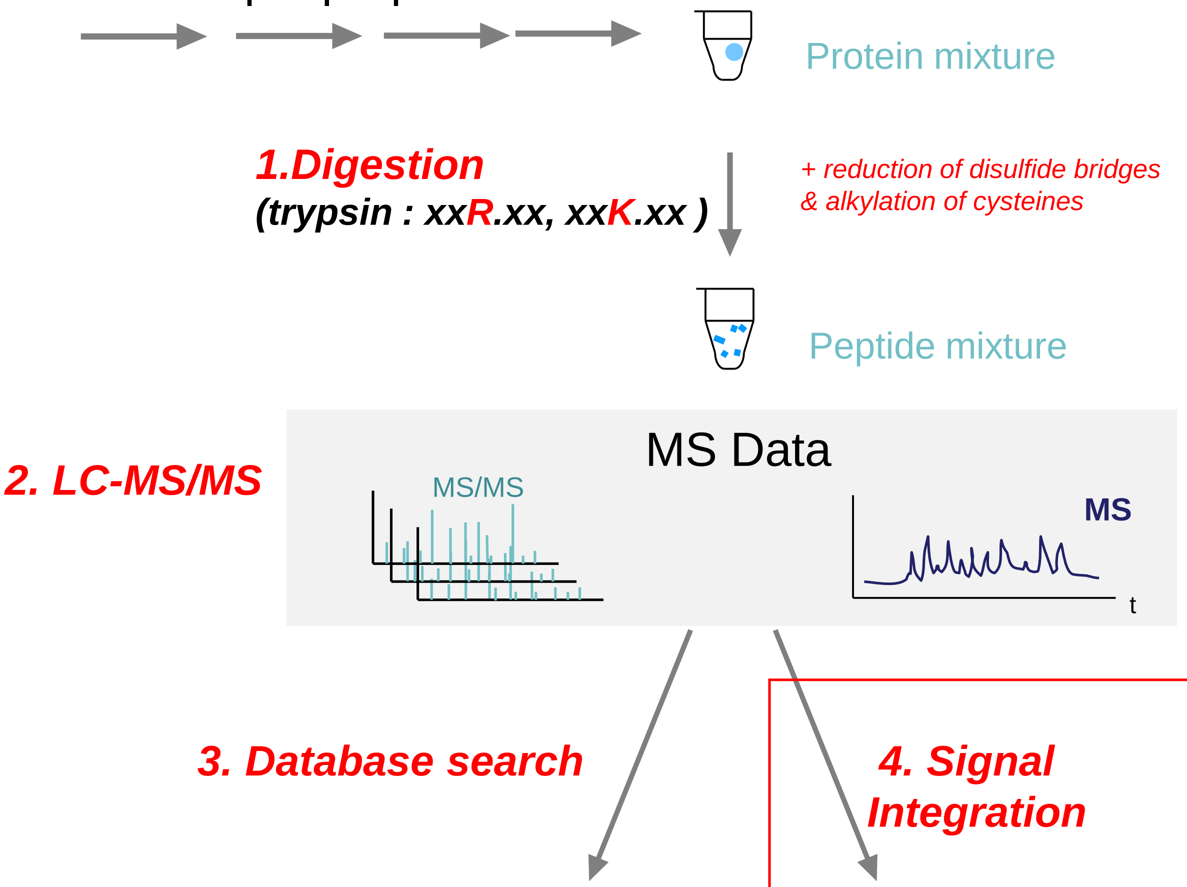
- **Protein levels : main end product of gene activation, functionally active molecules**

- **Transcriptomics (cDNA, Affymetrix oligo chips, RNAseq,…) vs. proteomics**
  - **Comprehensive**
  - **Higher throughput, fast(er)**
  - **More sensitive**
  - **Assumption : [mRNA] ~ [protein]**



Schwanhäusser B. *et al. Nature* **473**, 337-342 (2011)

Main pipeline (bottom-up proteomics)



Sample preparation

Protein mixture

1.Digestion

(trypsin : xxR.xx, xxK.xx)

+ reduction of disulfide bridges
& alkylation of cysteines

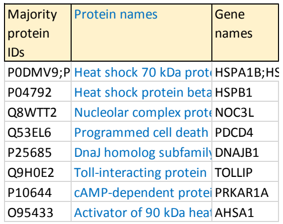
Peptide mixture

2. LC-MS/MS

MS Data

3. Database search

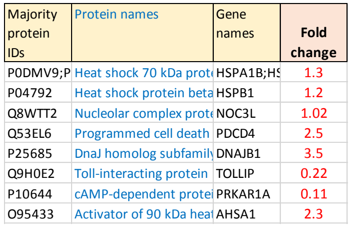
4. Signal Integration



Majority protein IDs	Protein names	Gene names
P0DMV9;P	Heat shock 70 kDa prot	HSPA1B;HS
P04792	Heat shock protein beta	HSPB1
Q8WTT2	Nucleolar complex prot	NOC3L
Q53EL6	Programmed cell death	PDCD4
P25685	DnaJ homolog subfamily	DNAJB1
Q9H0E2	Toll-interacting protein	TOLLIP
P10644	cAMP-dependent protei	PRKAR1A
O95433	Activator of 90 kDa heat	AHSA1

Protein ID

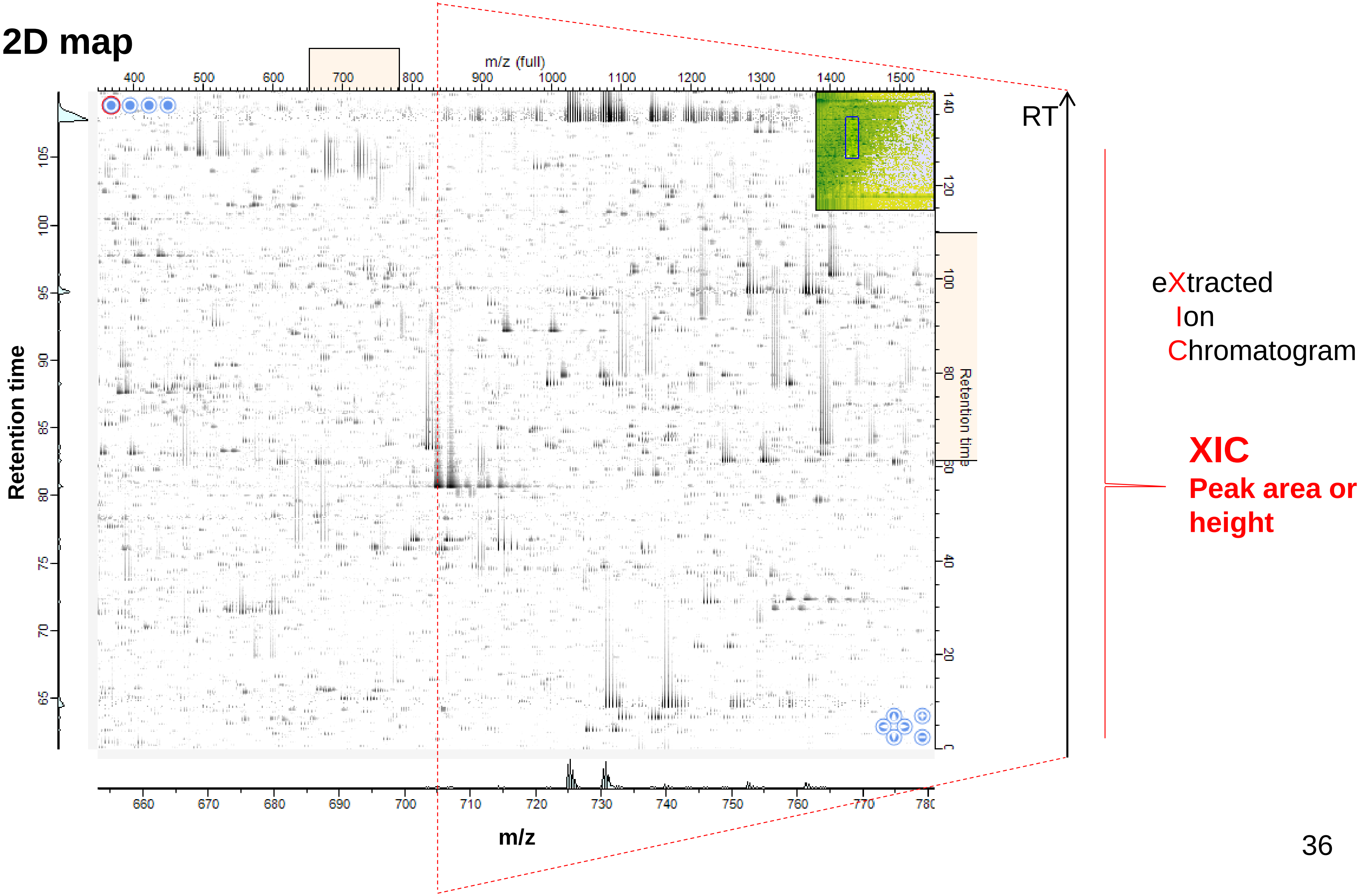
Protein quantitation



Majority protein IDs	Protein names	Gene names	**Fold change**
P0DMV9;P	Heat shock 70 kDa prot	HSPA1B;HS	1.3
P04792	Heat shock protein beta	HSPB1	1.2
Q8WTT2	Nucleolar complex prot	NOC3L	1.02
Q53EL6	Programmed cell death	PDCD4	2.5
P25685	DnaJ homolog subfamily	DNAJB1	3.5
Q9H0E2	Toll-interacting protein	TOLLIP	0.22
P10644	cAMP-dependent protei	PRKAR1A	0.11
O95433	Activator of 90 kDa heat	AHSA1	2.3

The XIC is used for quantitating peptide signals

2D map



eXtracted
Ion
Chromatogram

XIC

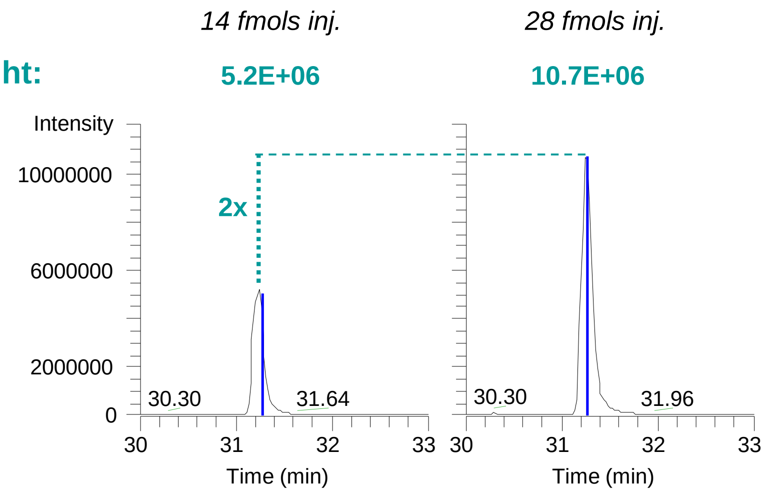
Peak area or height

Spectral counting *vs* MS1 XIC

Peptide **LVNELTEFAK** (BSA), *m/z* 582.320 (z=2)

Peak height:

Spectrum counting: **1**



Spectrum counting: **1**

- Spectral counting may not reflect actual intensity differences, especially for low signals
- Even when it does indicate a difference it is often not linear/accurate (stochasticity of precursor picking)
- **Better quantification with MS1 (XIC) signal**

Relative protein quantification

Comparison : A ⇔ B

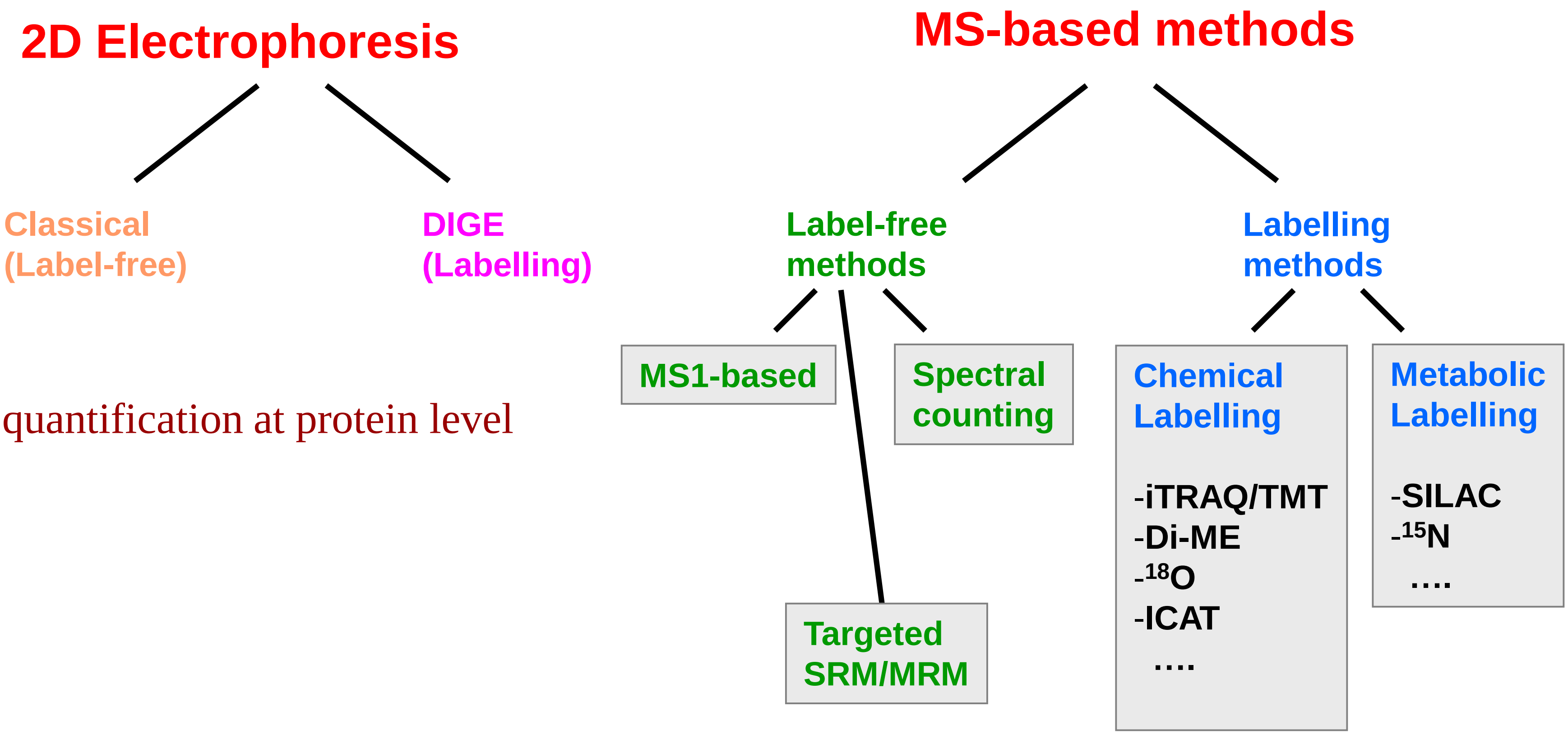
? Which proteins change in amount and how much ?

Applications :

- **Healthy vs. diseased tissues**
- **Healthy vs. diseased body fluids**
- **Drug treated / untreated cells**
- **Stimulated / unstimulated cells**
- **Mutants / wt cells**

…….

Techniques for large scale quantitative proteomics



2D Electrophoresis

Classical (Label-free)

DIGE (Labelling)

MS-based methods

Label-free methods

MS1-based

Spectral counting

Targeted SRM/MRM

Labelling methods

Chemical Labelling

-iTRAQ/TMT
-Di-ME
-^{18}O
-ICAT
….

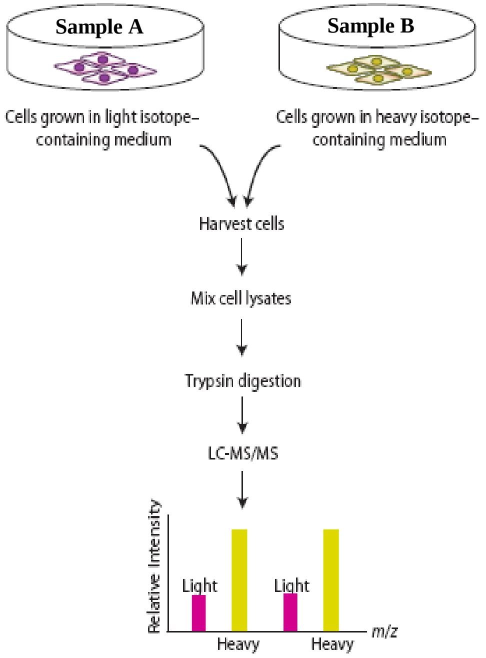
Metabolic Labelling

-SILAC
-^{15}N
….

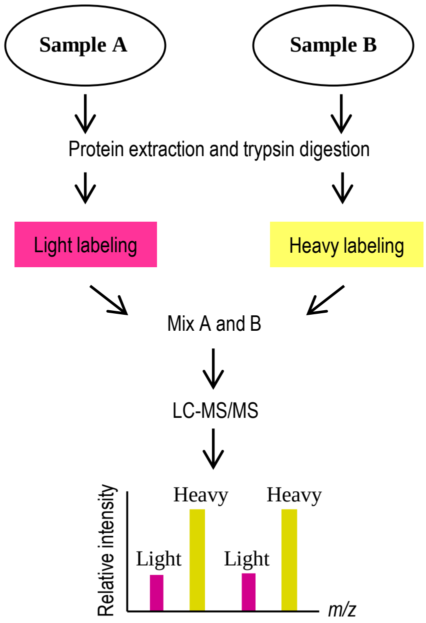
- ➢ quantification at protein level

- ➢ quantification at peptide level

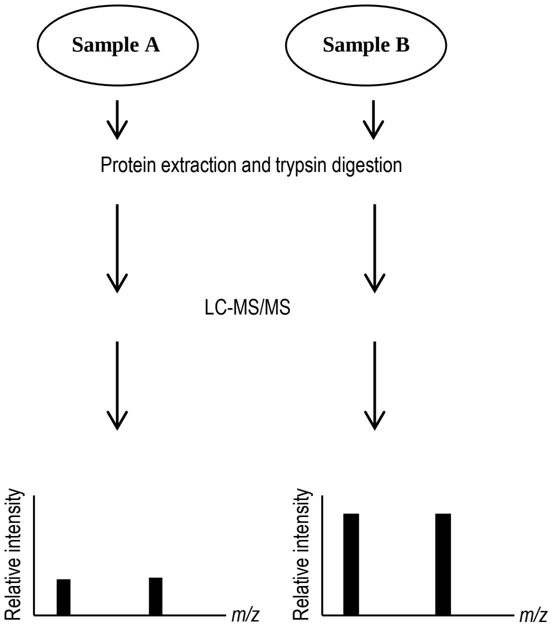
Metabolic labelling



Chemical labelling



Label free



Labeling

👍 Analytical variability minimized

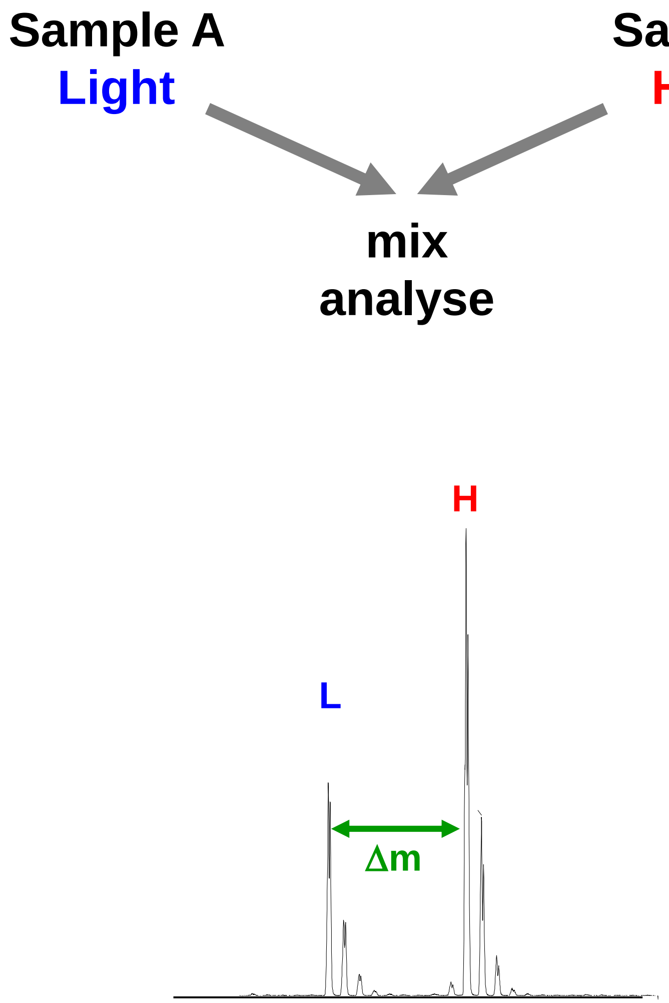
👎 Number of samples limited (2-8)

Label free

👍 Number of samples unlimited
Simpler sample preparation

👎 Analytical variability
<u>Computationally heavy (XIC)</u>

Relative quantification by stable isotope labelling


Sample A
Light

Sample B
Heavy

mix
analyse

Labelling strategies :

- Chemical (side chains : C, K, N-term)
  ICAT, iTRAQ, ICPLP,…
- Metabolic (K, R, all)
  SILAC,…
- Enzymatic
  Trypsin + 18O, …

Co-analyse

⇩

Eliminate analytical variability

How to label ? Pros and cons

- **Metabolically** (during protein synthesis)

→ Incorporation of one or more labelled amino acid

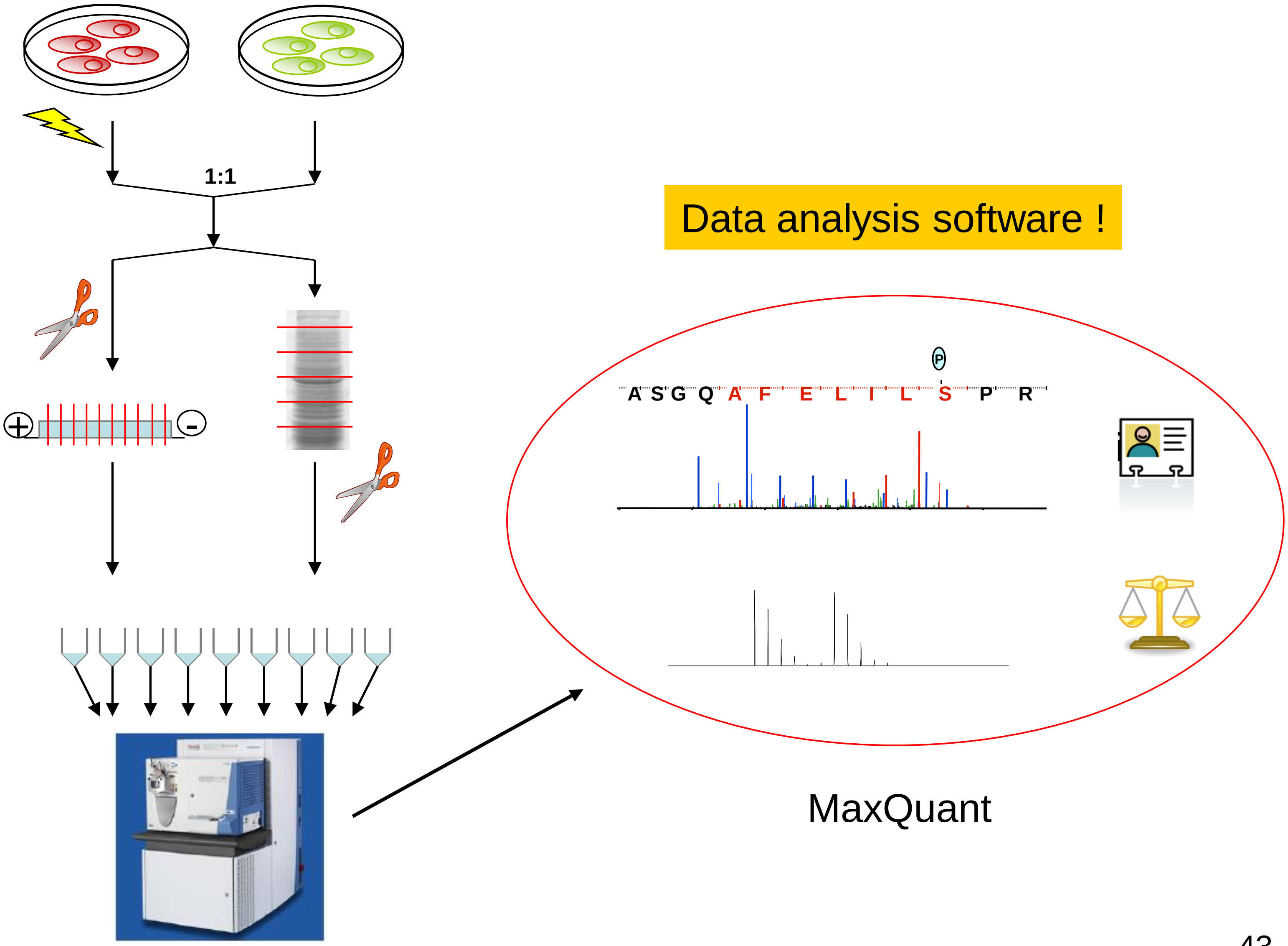
(+) “native” proteins
(+) compatible w. purifications
(+) accurate
(-) need cultivatable organism
(-) limited multiplexing (max. 3)

- **Chemically** (post protein synthesis)

→ “specific” chemical modification of AA side chain

(+) any sample can be done
(+) higher multiplex (iTRAQ max 8-plex)
(-) side (or incomplete) reactions
(-) separate purifications
(-) less accurate

SILAC experiment workflow

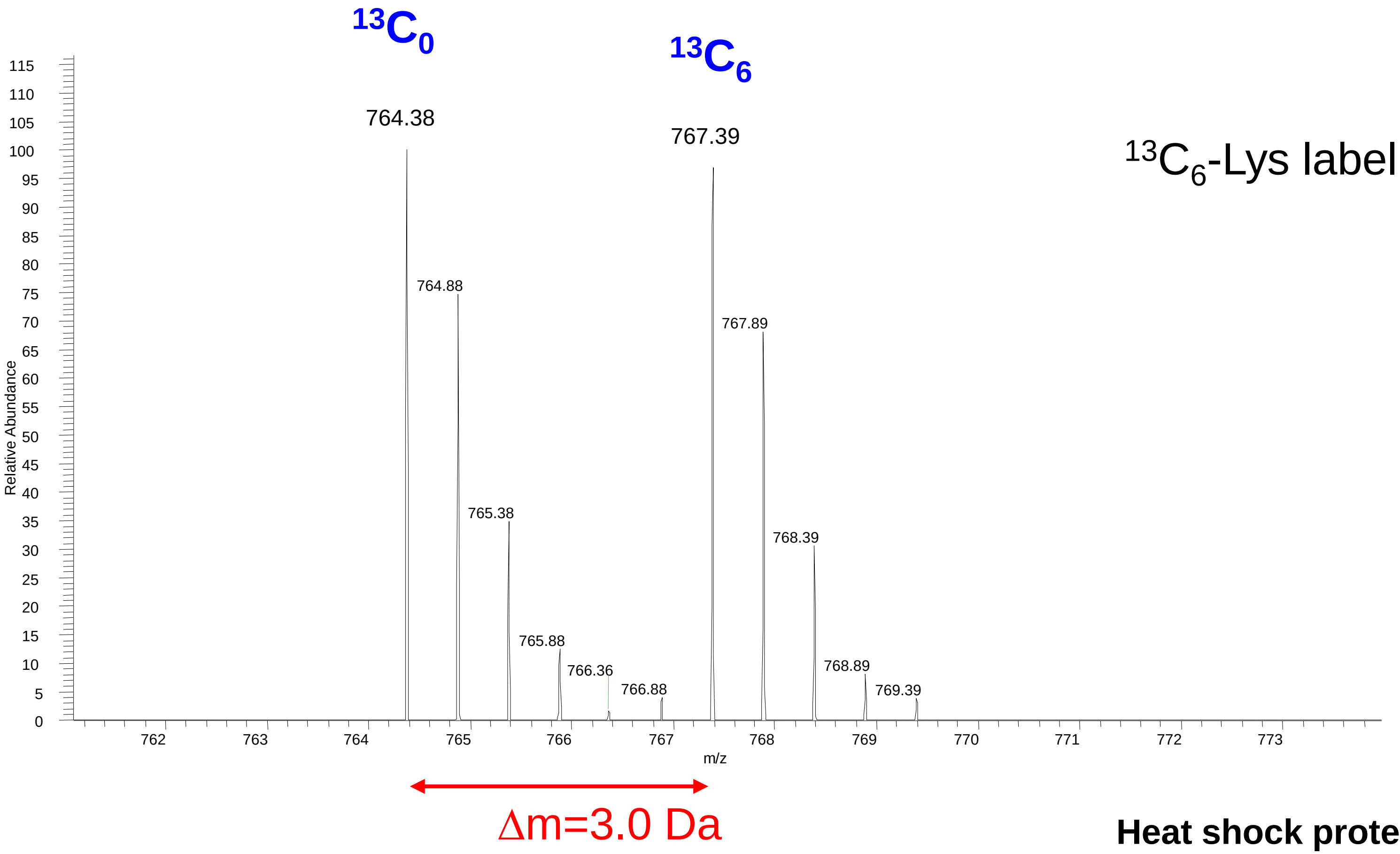


Data analysis software !

MaxQuant

SILAC peaks

…K SLTNDWEDHLAVK H…



$^{13}C_6$-Lys labelling

Heat shock protein
HSP 90β

Chemical labelling :
Isobaric Tags (iTRAQ)- multiplex quantification

Multiplexed Protein Quantitation in *Saccharomyces cerevisiae* Using **Amine-reactive Isobaric Tagging Reagents**
Ross P., Pappin D. et al. *Molecular & Cellular Proteomics 3.12 - 2004*

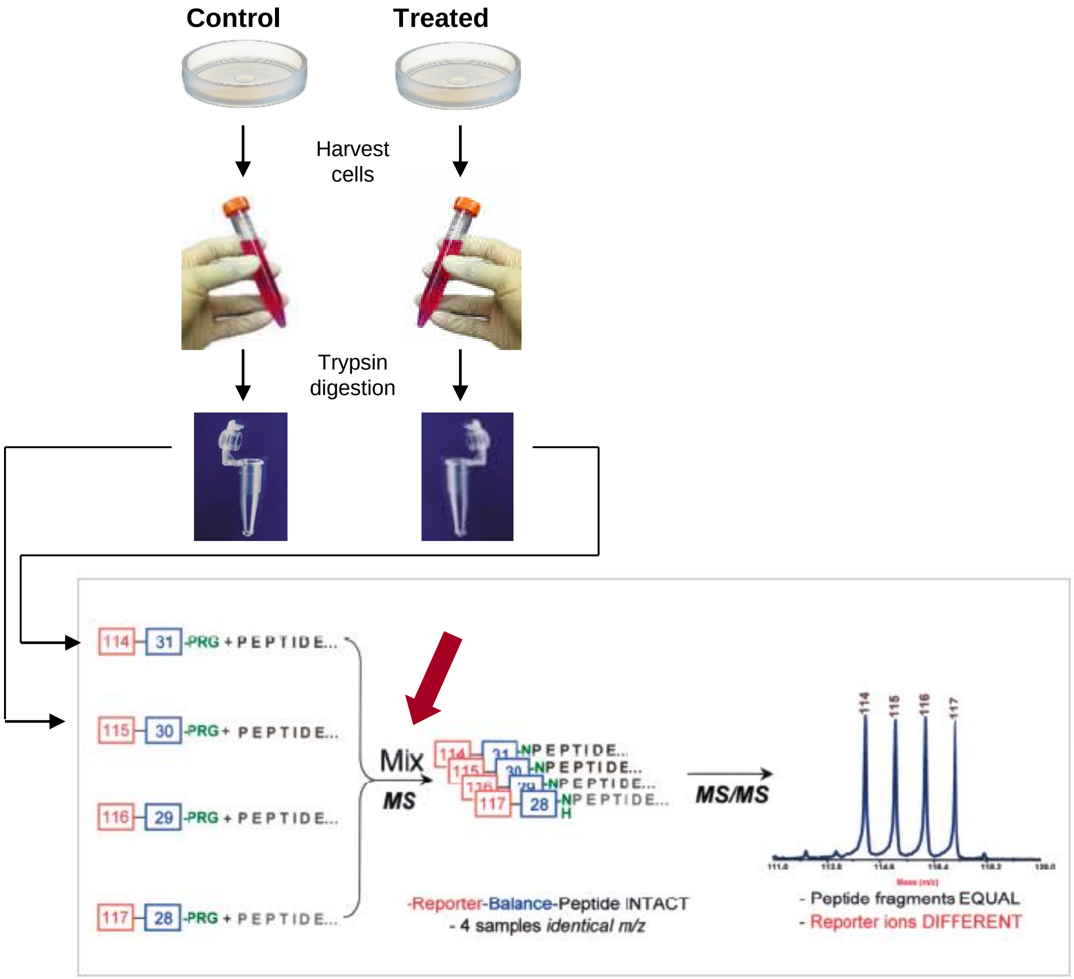
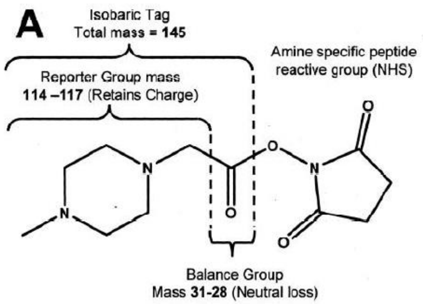


Figure 1. The concept of iTRAQ™ Reagent chemistry (example of a 4-plex experiment) Each sample is labeled with one of the four iTRAQ Reagents and then pooled prior to MS analysis.

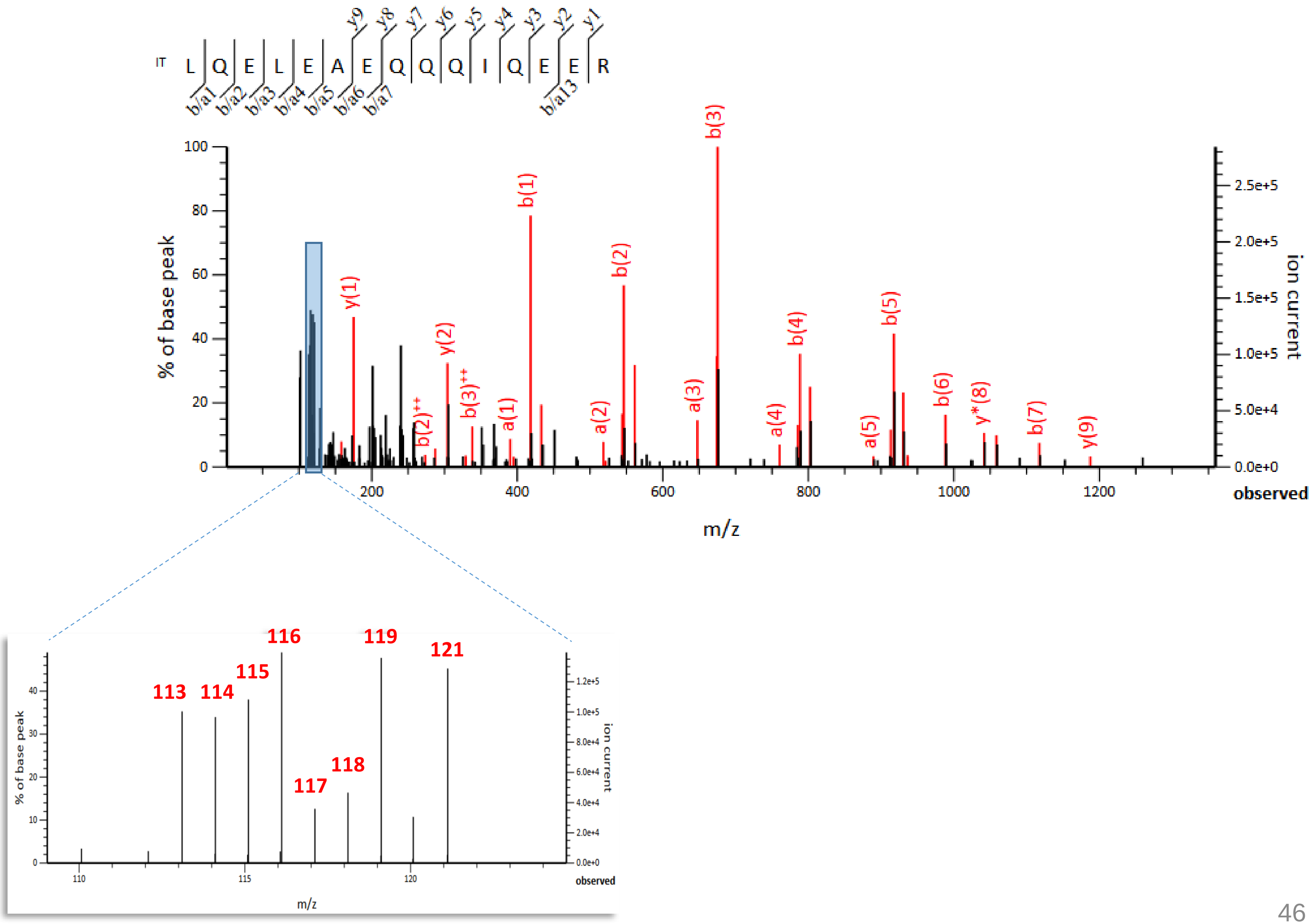

A
Isobaric Tag
Total mass = 145
Amine specific peptide reactive group (NHS)
Reporter Group mass
114 –117 (Retains Charge)
N
N
O
O
N
O
O
Balance Group
Mass 31-28 (Neutral loss)

← ***in vitro* chemical labeling**

← **Same peptide from 4 samples has same mass (isobaric)**

← **quantification by tags in MS/MS spectra at fixed M/Z**

Information Content of iTRAQ MS2 spectra



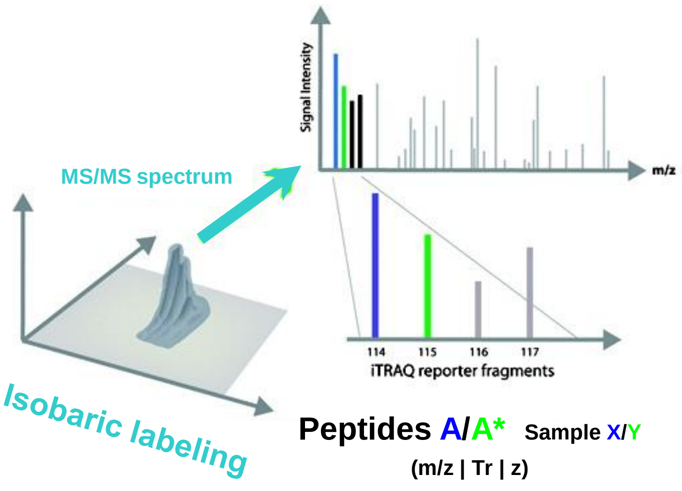
iTRAQ/TMT formats and reagents

	Multiplexing	Target functional groups	Mass range (max)	Features
iTRAQ	4-, 8-plex	NH2 (N-term, Lys)	113-121	General peptide labelling
TMT	2-, 6-, 10-, 11-plex	NH2 (N-term, Lys)	126-131	General peptide labelling ; 10, 11-plex need higher scan resolution to separate reporter ions
TMT	6-plex	-SH (Cys)		Selective for Cysteines
TMT	6-plex	C=O		Glycans, steroids, oxidized proteins

McAlister, G. C., Huttlin, E. L., Haas, W., Ting, L., Jedrychowski, M. P., Rogers, J. C., … Gygi, S. P. (2012). Increasing the multiplexing capacity of TMTs using reporter ion isotopologues with isobaric masses. *Analytical Chemistry*, *84*(17), 7469–78. http://doi.org/10.1021/ac301572t

Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., … Pappin, D. J. (2004). Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Molecular & Cellular Proteomics : MCP*, *3*(12), 1154–69. http://doi.org/10.1074/mcp.M400129-MCP200

Signal processing in MS quantification

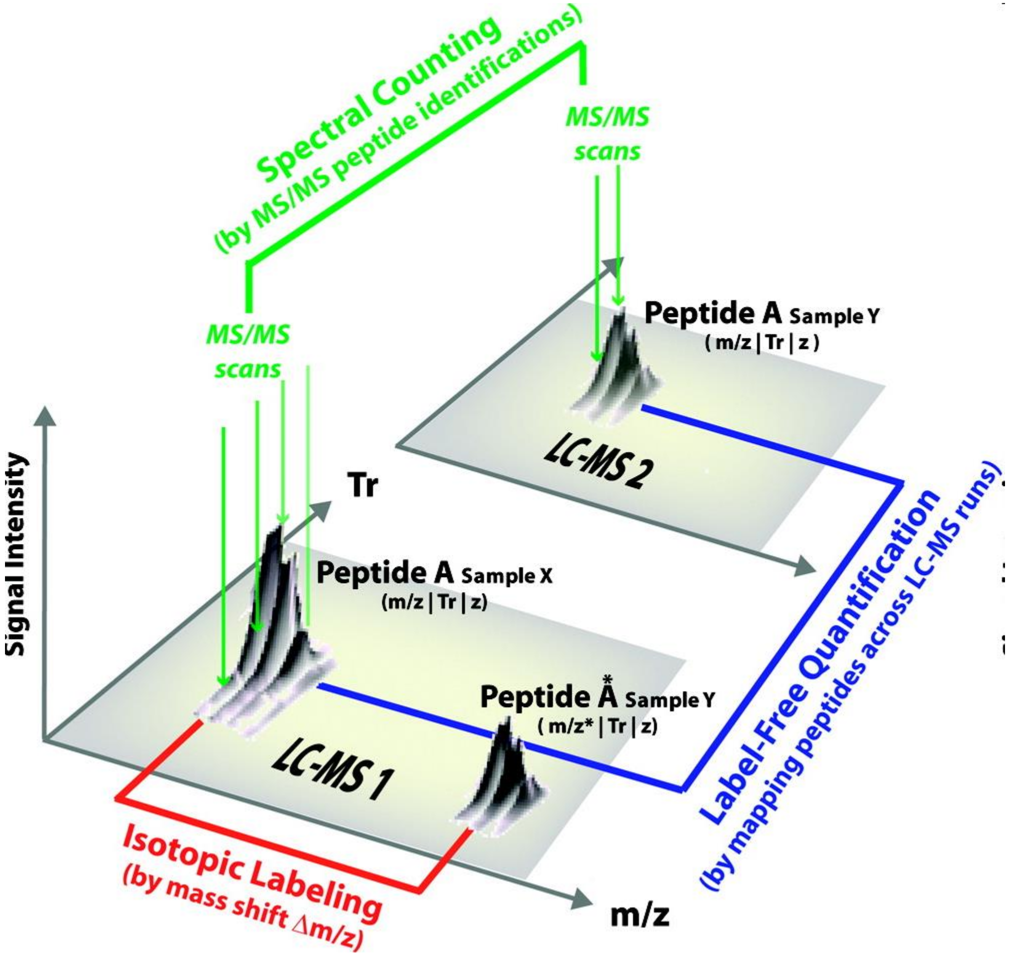


MS/MS spectrum

Isobaric labeling

Peptides A/A* Sample X/Y

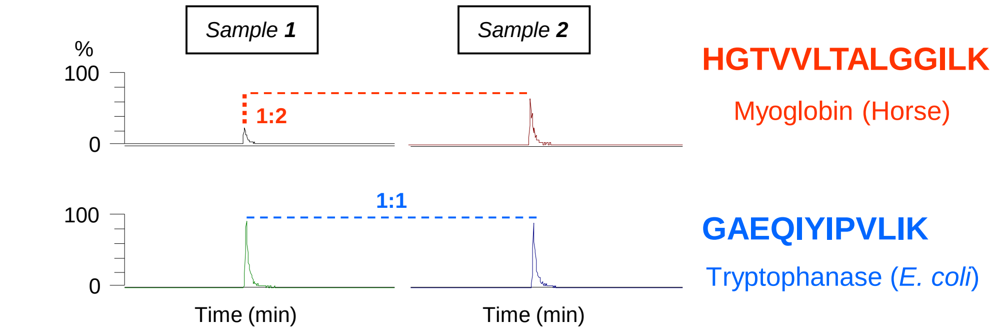
(m/z | Tr | z)



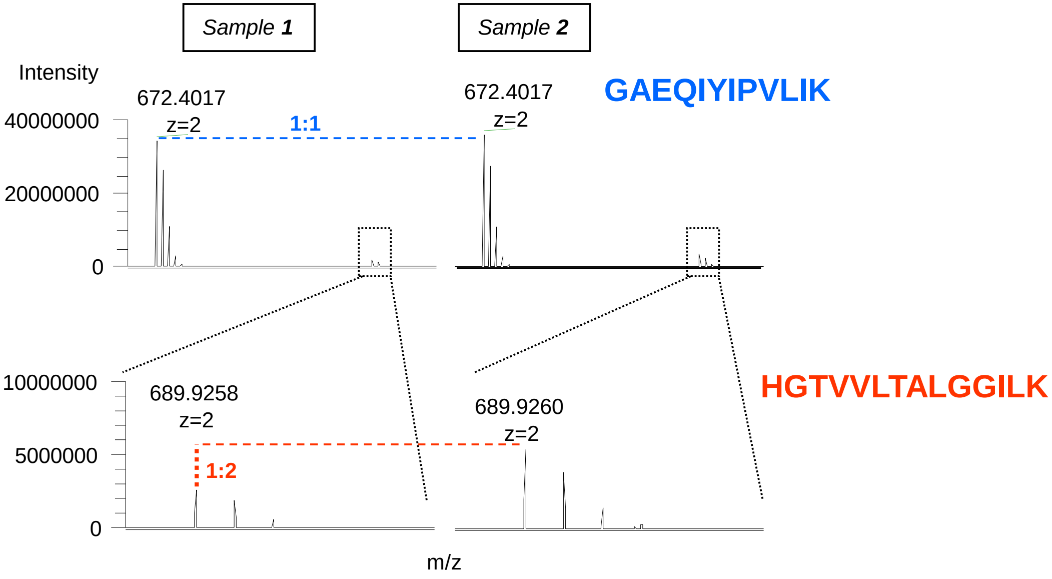
Figures from: *Mueller L.N., Brusniak M.Y., Mani D.R., Aebersold R. Journal of Proteome Research, 2008, 7(1), 51-61.*

Example of label-free quantitation

Spiked myoglobin (ratio 1:2) in *E.coli* lysate

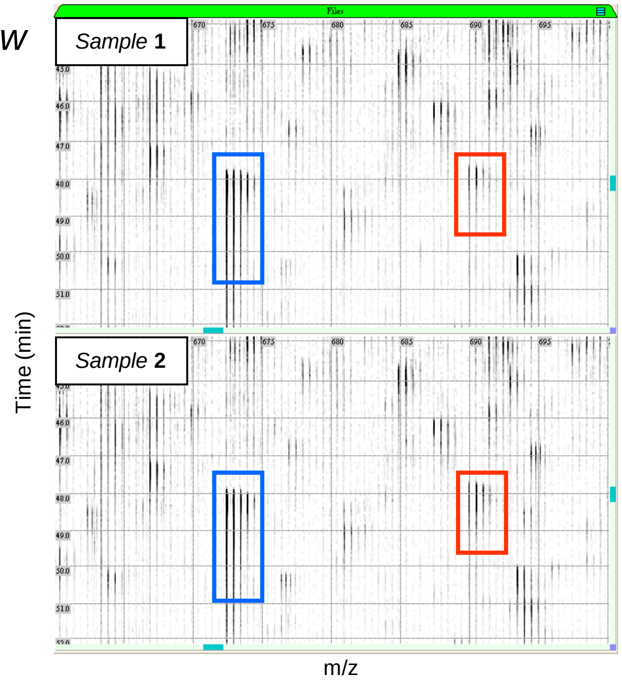


Chromatogram view



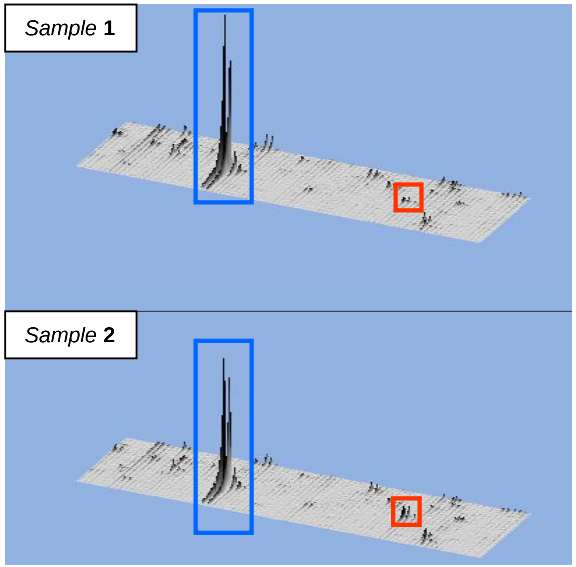
Spectrum view

2D view



MSight (http://web.expasy.org/MSight)

3D View



Processing in label-free quantitation (XIC)

Main Steps:

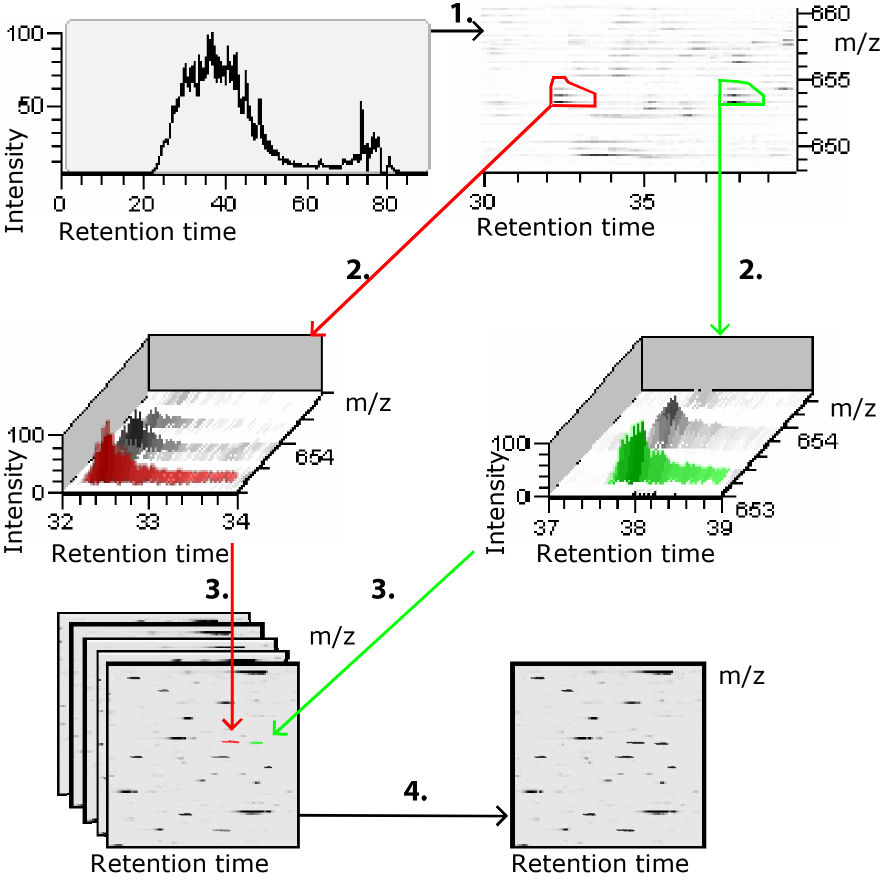
Preprocessing

Feature (m/z, tr, z, intensity) extraction

Matching and alignment

Normalisation

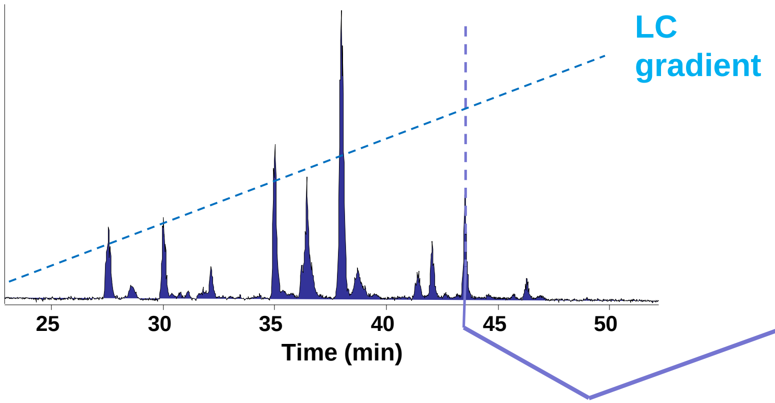
Statistical Analysis

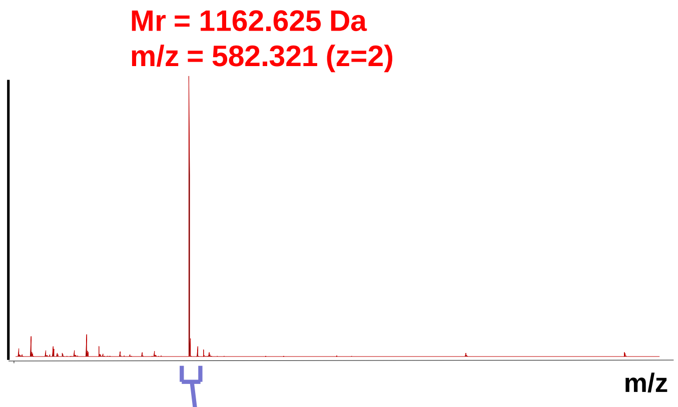


1. feature detection in each LC-MS run
2. quantification: intensity integration
3. creation of 2D feature map
4. matching and alignment across LC-MS runs

LC-MS/MS data dependent acquisition (DDA)

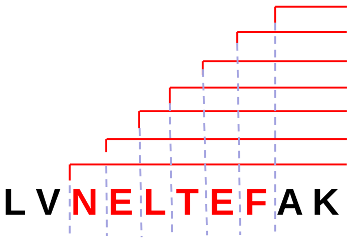
1. MS scan: Determination of peptide mass(es)

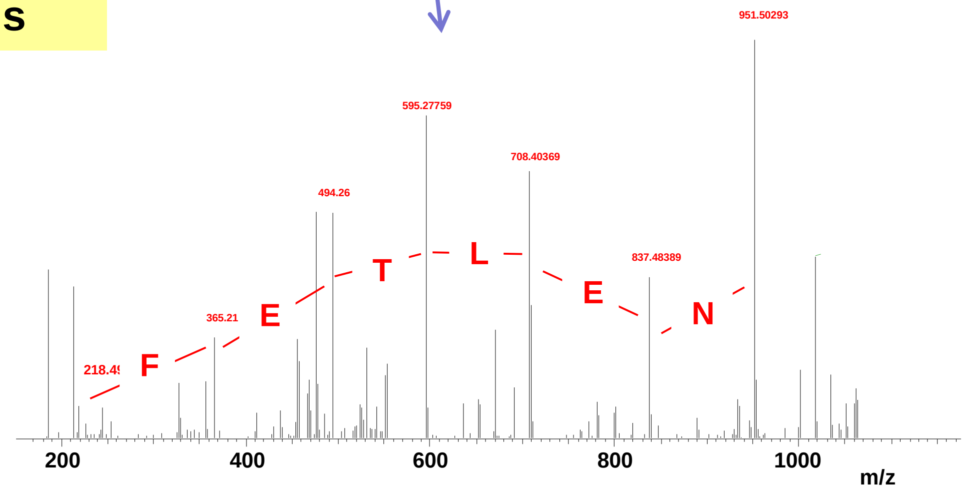



Mr = 1162.625 Da
m/z = 582.321 (z=2)

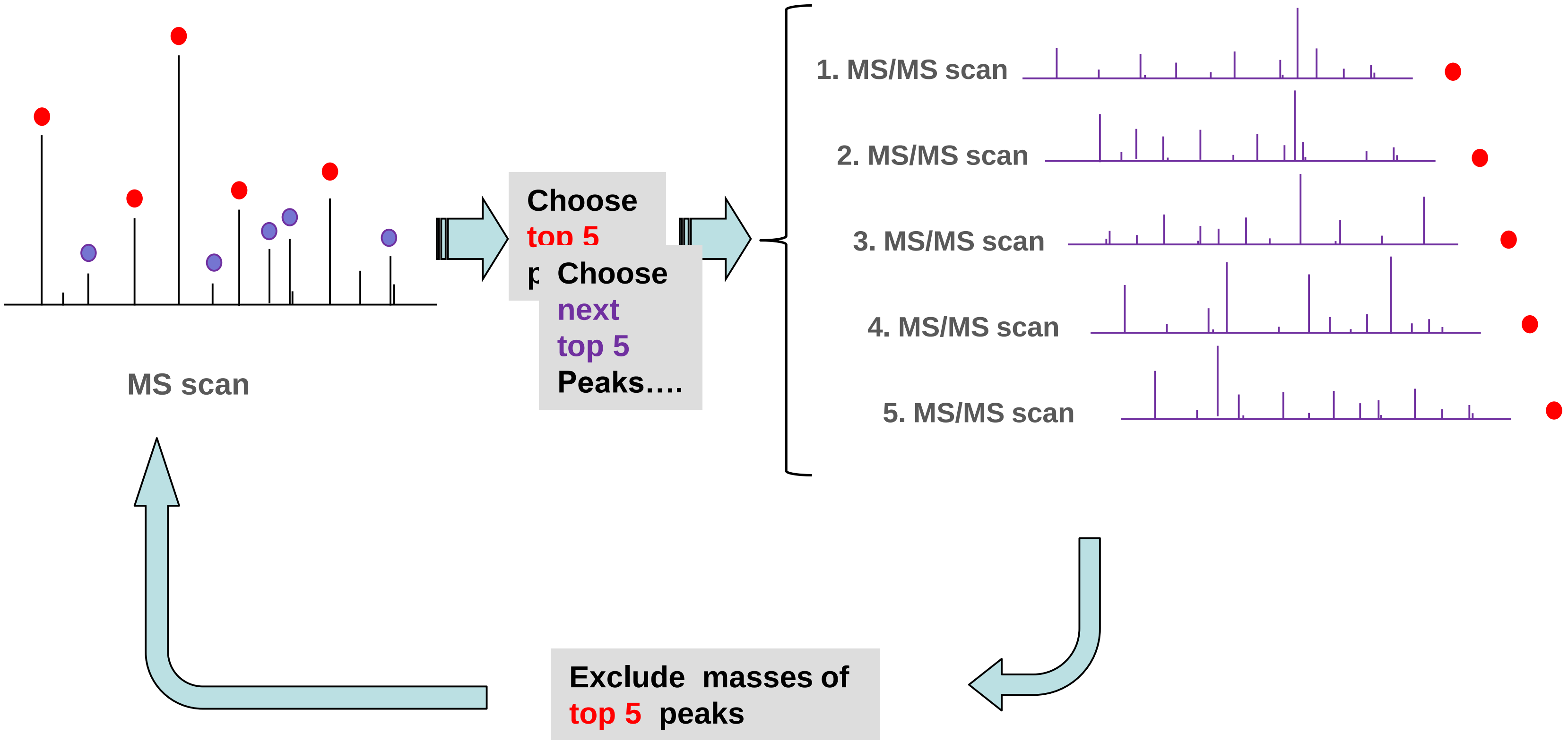
Ion isolation
Collision Induced Dissociation (CID)

2. MS/MS scan: Isolation, fragmentation, fragment analysis





Automated DDA cycle



MS scan

Choose
top 5
p

Choose
next
top 5
Peaks….

1. MS/MS scan

2. MS/MS scan

3. MS/MS scan

4. MS/MS scan

5. MS/MS scan

Exclude masses of
top 5 peaks

Goal : selection and fragmentation of a maximum number of peptides per unit of time

Missing data problem in DDA

- Data Dependent Acquisition (DDA) is a highly flexible method that can deal with the most diverse samples
- However, precursor selection in DDA is partially stochastic (or better, non deterministic) since it is based on contingent factors
- Especially in complex samples, the set of chosen (low abundance) precursors varies, even between replicate injections of the same sample

→ missing peptide ID's across samples (even when precursor is present)

→ Loss of significant numbers of peptides/proteins values

repl.1	repl.2	repl.3
23.1	21.3	22.5
18.7	17.2	14.3
8.9	9.9	NaN
26.2	25.3	23.4
20.1	25.2	19.4
7.3	6.4	8
NaN	21.2	17.2
14.2	NaN	15.3
18.1	18.2	19.6
11.2	NaN	13.2

Solutions :

1) Match between runs (MaxQuant)
2) **<u>D</u>ata-<u>I</u>ndependent <u>A</u>cquisition**

Quantitation tools : choice is limited

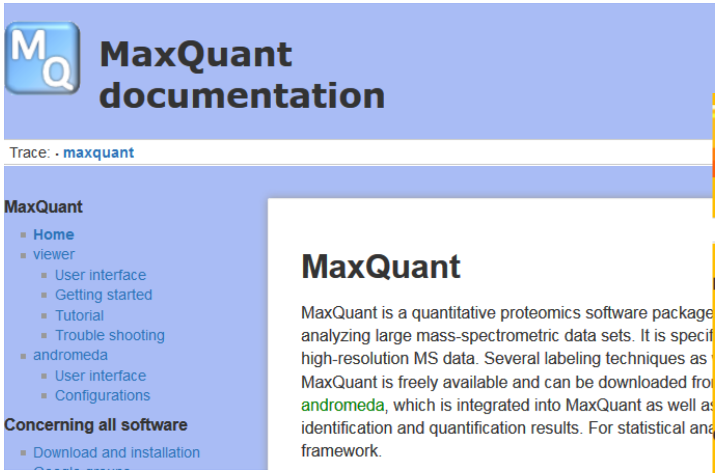
- Proteome Discoverer (Thermo Fisher Scientific)
  - Good GUI
  - Workflows and working with user plugins, other search engines => versatility
  - Compatibility only w. Thermo data
  - Commercial => expensive, licenses limit processing usage
  - Windows only

- MaxQuant/Perseus :
  - Freeware (but not open source) academic product
  - Versatile : LFQ, SILAC, iTRAQ, …
  - Complex experimental designs possible
  - Extensive data output tables
  - Scalable (100’s, 1000’s of files)
  - Coherent, long term development
  - Windows mainly but LINUX coming next
  - No manual (though web resources available)=> parameters ?
  - No release notes

- Many other commercial packages (none is of broad scope)
- OpenMS initiative…interesting but not well known
- Skyline (DDA, DIA, targeted quantification)
- Other *ad hoc* academic software tools : poor/difficult diffusion outside originating lab

Note : many other search engines exist but mostly not (or not well) coupled to quantitation

MaxQuant/Perseus environment



MaxQuant documentation

Trace: · maxquant

MaxQuant

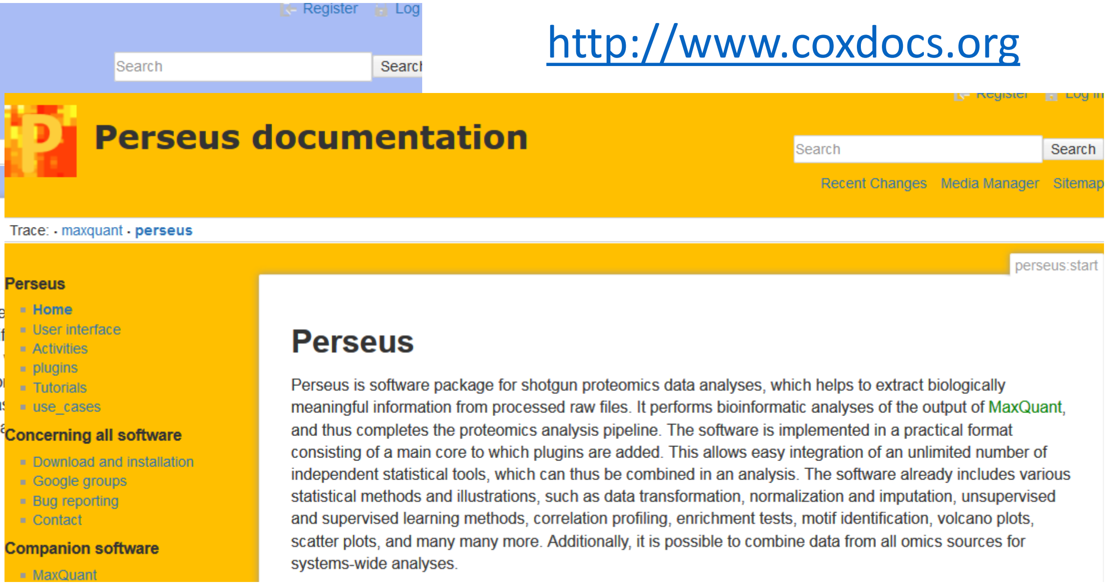
- Home
- viewer
  - User interface
  - Getting started
  - Tutorial
  - Trouble shooting
- andromeda
  - User interface
  - Configurations

Concerning all software

- Download and installation

MaxQuant

MaxQuant is a quantitative proteomics software package analyzing large mass-spectrometric data sets. It is specif high-resolution MS data. Several labeling techniques as MaxQuant is freely available and can be downloaded fro andromeda, which is integrated into MaxQuant as well a identification and quantification results. For statistical ana framework.



Perseus documentation

Trace: · maxquant · perseus

Perseus

- Home
- User interface
- Activities
- plugins
- Tutorials
- use_cases

Concerning all software

- Download and installation
- Google groups
- Bug reporting
- Contact

Companion software

- MaxQuant

Perseus

Perseus is software package for shotgun proteomics data analyses, which helps to extract biologically meaningful information from processed raw files. It performs bioinformatic analyses of the output of MaxQuant, and thus completes the proteomics analysis pipeline. The software is implemented in a practical format consisting of a main core to which plugins are added. This allows easy integration of an unlimited number of independent statistical tools, which can thus be combined in an analysis. The software already includes various statistical methods and illustrations, such as data transformation, normalization and imputation, unsupervised and supervised learning methods, correlation profiling, enrichment tests, motif identification, volcano plots, scatter plots, and many many more. Additionally, it is possible to combine data from all omics sources for systems-wide analyses.

http://www.coxdocs.org

Google groups :
https://groups.google.com/forum/#!forum/perseus-list

YouTube videos :
https://www.youtube.com/c/MaxQuantChannel

Workflows on Doku-Wiki pages http://www.coxdocs.org/doku.php?id=perseus:user:use_cases:start

MaxQuant refs

Cox, J. and Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol, 2008, 26, pp 1367-72.

Cox J., Hein M. Y., Luber C. A., Paron I., Nagaraj N., and Mann M., Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. Mol Cell Proteomics, 2014, 13, pp 2513–2526.

Cox, J., Matic, I., Hilger, M., Nagaraj, N., Selbach, M., Olsen, J. V, & Mann, M. (2009). A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics. Nature Protocols, 4, 698–705. http://doi.org/10.1038/nprot.2009.36

Perseus refs

Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., … Cox, J. (2016). The Perseus computational platform for comprehensive analysis of (prote)omics data. Nature Methods, 13(9), 731–40. http://doi.org/10.1038/nmeth.3901

Tyanova, S., & Cox, J. (2018). Perseus: A Bioinformatics Platform for Integrative Analysis of Proteomics Data in Cancer Research. Methods in Molecular Biology (Clifton, N.J.), 1711, 133–148. http://doi.org/10.1007/978-1-4939-7493-1_7

MaxQuant output variables (protein level)

SILAC

- **H/L ratio*** (raw / normalized)
- Intensities (H,L,Total)
- [iBAQ]

Most accurate for relative quant.
Includes Re-quantify step

Measurement different from H/L

Only for global abundance

Label-free

- Intensity
- **LFQ***
- [iBAQ]

LFQ most accurate IF normalization is feasible
Can include *match between runs* function

Once normalized, can be used for absolute quant and copy numbers

iTRAQ/TMT

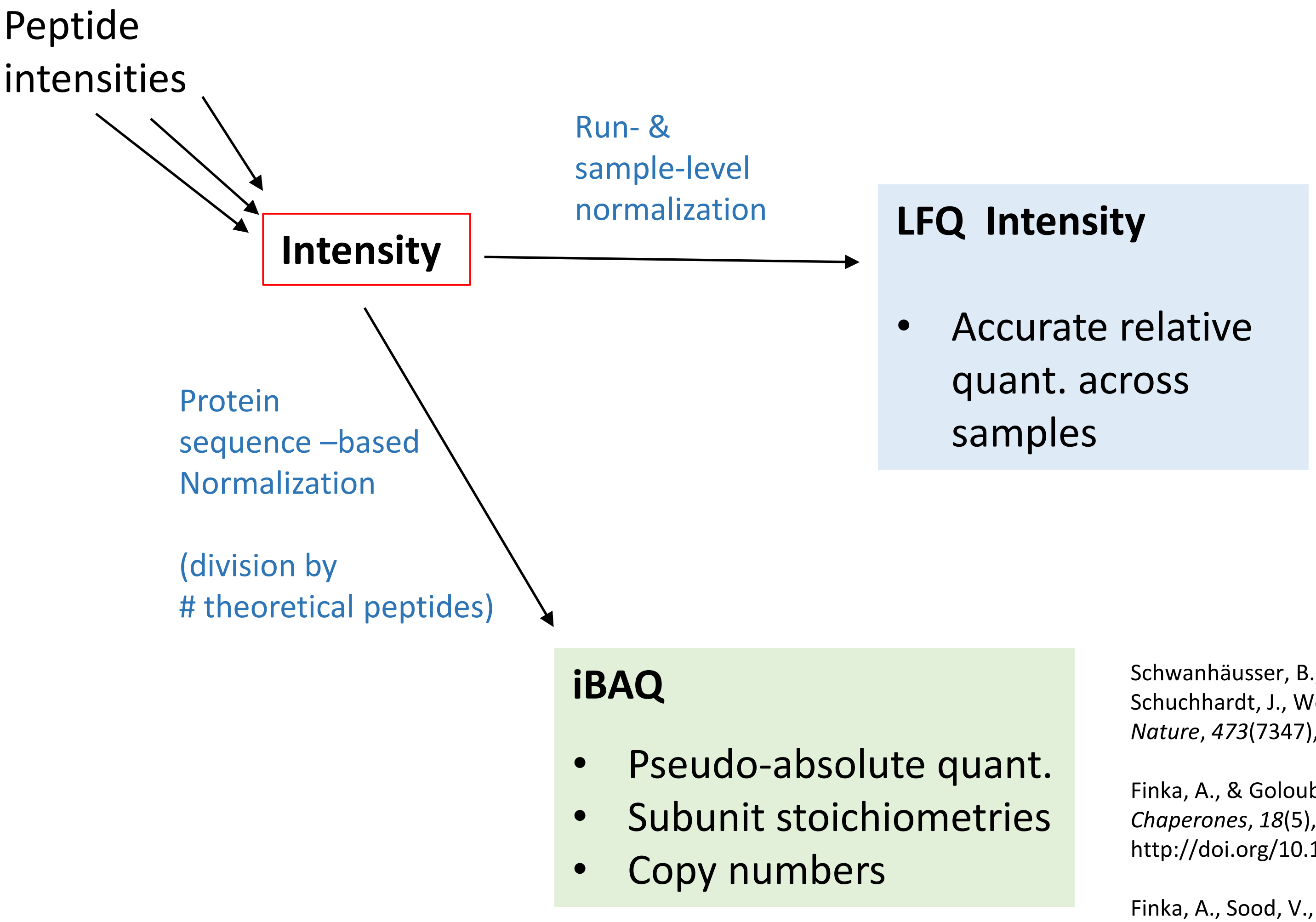
- **Reporter intensities***
- Intensities
- [iBAQ]

Sum of all peptide RI intensities is used !

Only (maybe…) for global abundance

*** : main output**

MaxQuant label-free output variables (protein level)



Peptide intensities

Intensity

Run- & sample-level normalization

LFQ Intensity

- Accurate relative quant. across samples

Protein sequence –based Normalization

(division by # theoretical peptides)

iBAQ

- Pseudo-absolute quant.
- Subunit stoichiometries
- Copy numbers

Cox, J., Hein, M. Y., Luber, C. a, Paron, I., Nagaraj, N., & Mann, M. (2014). *Molecular & Cellular Proteomics : MCP*, *13*(9), 2513–26.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., … Selbach, M. (2011). *Nature*, *473*(7347), 337–42.

Finka, A., & Goloubinoff, P. (2013). *Cell Stress & Chaperones*, *18*(5), 591–605. http://doi.org/10.1007/s12192-013-0413-3

Finka, A., Sood, V., Quadroni, M., Rios, P. D. L., & Goloubinoff, P. (2015). *Cell Stress and Chaperones*, 605–620.

Quantitation summary

	Application	Multiplexing	Accuracy (process)	Quantitative proteome coverage	Linear dynamic range[a]	Ease of use
Metabolic protein labeling	• Complex biochemical workflows • Cell culture systems only	2-3	+++	++	1–2 logs	+
Chemical protein labeling (MS)	• Medium to complex biochemical workflows	2-3	+++	++	1–2 logs	+
Chemical peptide labeling (MS)	• Medium complexity biochemical workflows	2-3	++	++	2 logs	+
Chemical peptide labeling (MS/MS)	• Medium complexity biochemical workflows	2-8	++	++	2 logs	+
Enzymatic labeling (MS)	• Medium complexity biochemical workflows	2	++	++	1–2 logs	++
Spiked peptides	• Medium complexity biochemical workflows • **Targeted analysis of few proteins**	multiple	++	+	2 logs	++
Label free (ion intensity)	• Simple biochemical workflows • Whole proteome analysis	multiple	+	+++	2–3 logs	++
Label free (spectrum counting)	• Simple biochemical workflows • Whole proteome analysis	multiple	+	+++	2–3 logs	+++

[a] In MRM mode, dynamic range may be extended to 4–5 logs

Adapted from:
Bantscheff M. et al. Anal Bioanal Chem (2007) 389:1017–1031.

Data processing in quantitative proteomics

- Experimental design: think about statistical testing before experiments ! Better to discuss with a biostatistician to pick up the correct test (power, data independence, parametric/non-parametric, hypothesis to test, ...):

"To call in the statistician after the experiment is done may be no more than asking him to perform a postmortem examination: he may be able to say what the experiment died of."
Ronald Fisher, Indian Statistical Congress, Sankhya, around 1938.

•From peptide to protein quantitation:

→ **median (resistant to outliers), min. 3 values (peptides, "evidence" or spectra values ?!)**

- Normalization of data according to some assumption:

→ **ex: most proteins don't change**

- Dealing with ratios: log-transformation (usually $\log_2$):

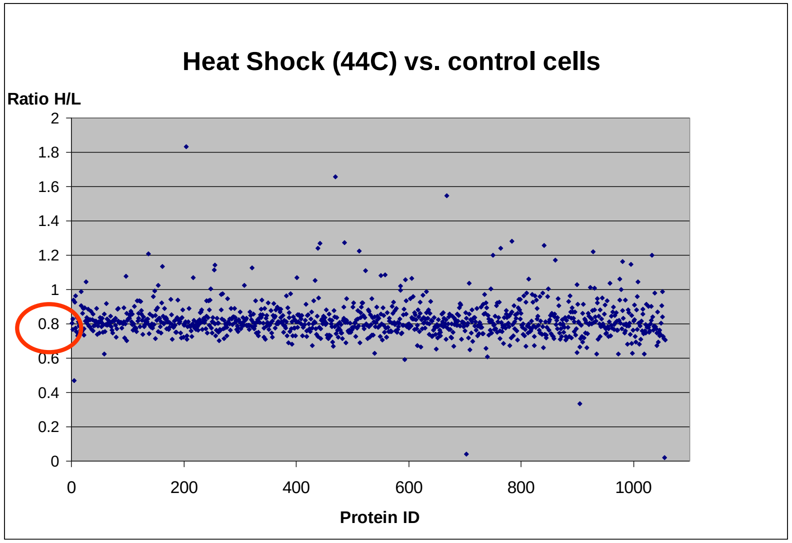
→ **data symmetric and "more" normal → statistics**

- Exploratory data analysis (descriptive statistics):

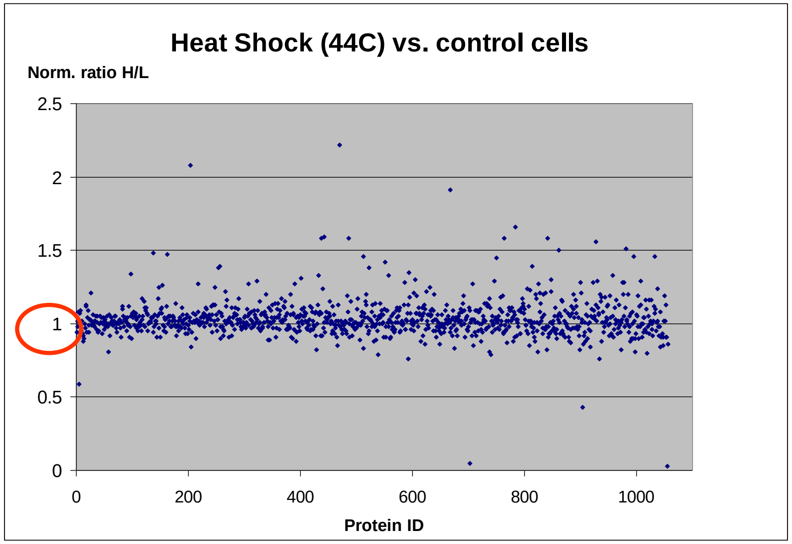
→ **numerical and graphical summaries**

Quantitation exercise

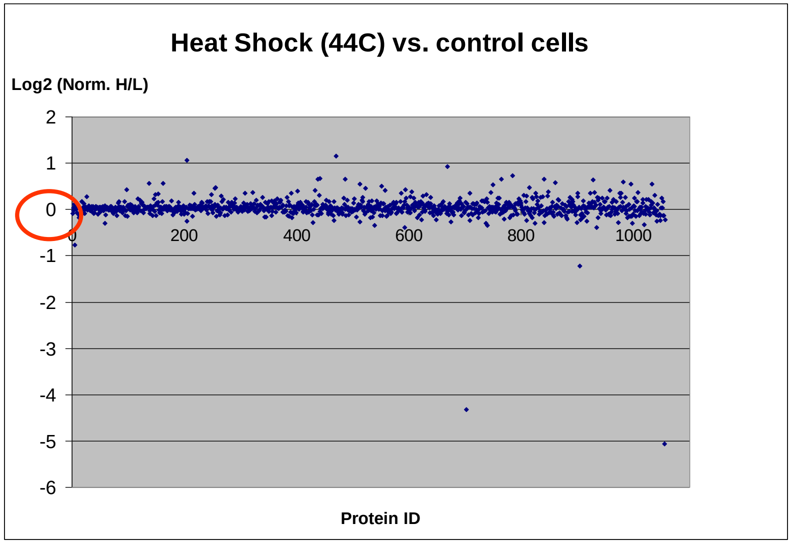
Data processing: summary



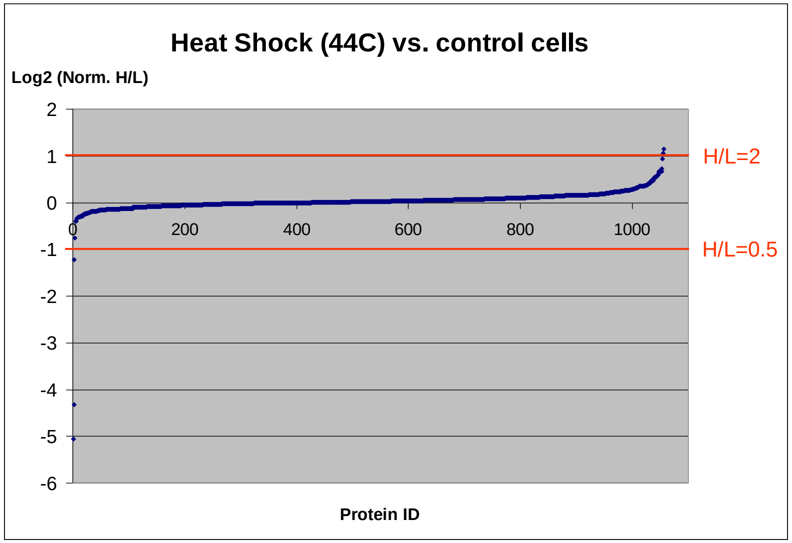
Normalization



Log_2 transformation



Sorting by ratio



Statistics & validation - summary

- Missing data problem

→ Imputation (to use with caution !)

- Multiple testing problem: false positives

→ FPR (Bonferroni, stringent) or FDR (Benjamini & Hochberg, less conservative) correction

- Significant differences ≠ meaningful differences

→ Statistical significance does not mean biological significance: minimum fold change threshold

- Statistical criteria stringency will depend on downstream data analysis

→ Is the aim of analysis a confident list of varying proteins or an overview of the proteome dynamics ?

Publication guidelines (MCP): quantitative results

- **Experiments**
    - How the quantitation was performed (number of peaks, peak intensity peak area, XIC)
    - Minimum thresholds required for data to be used for quantitation
    - Justification of removal of outlier data points
    - Explanation of statistics used to assess accuracy and significance of measurements
    - Indication of how biological and analytical reproducibility was addressed by experimental design

    => Biological replicates are almost mandatory these days…!

- **Results**
    - Number of peptides used for protein quantitation measurement
    - Protein quantitation measurement and accuracy (e.g. mean and standard deviation).

→ see also guidelines for reporting protein identification and PTMs

http://www.mcponline.org/site/misc/MSDataResources.xhtml

Targeted quantification

- Shotgun MS/MS : « fishing » experiment; sometimes desired molecule(s) are not detected

↓

- Need targeted techniques : measure what we want (*ex. Western Blot*)
- Need to increase robustness, obtain absolute (not relative) quantification
- Expected : more sensitivity through selectivity

Targeted quantification by Selected Reaction Monitoring (SRM)

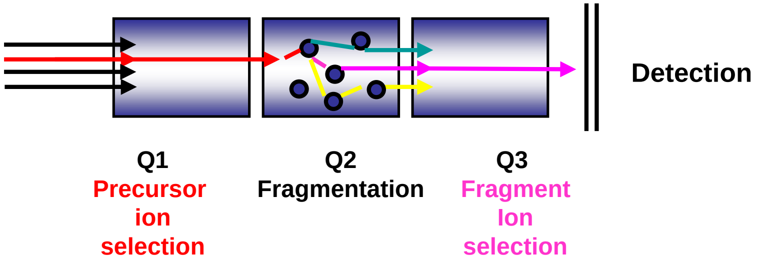
- **Follows proteomics discovery phase**
- **Targeted quantitation of proteins through « proteotypic » peptides**
- **Simplified MS / MS**
- **Any peptide (+PTM) can be measured**
- **Absolute quantitation (if done w. synthetic internal labeled standard)**
- **A few hundreds proteins measured in few hours**

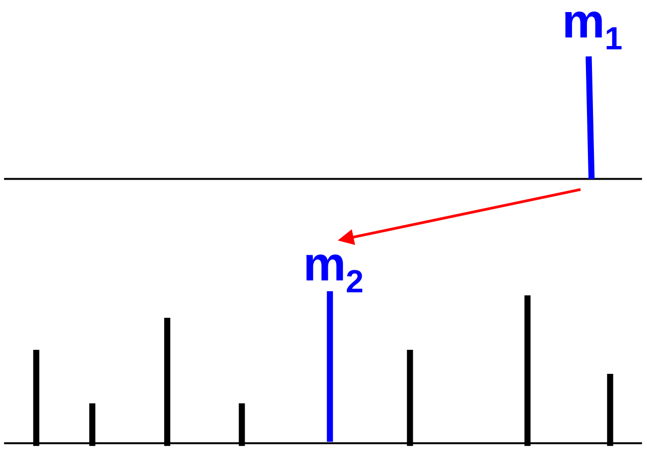
```
KPYM_HUMAN , Pyruvate kinase isozymes M1/M2 - Homo sapiens

1   MSKPHSEAGT AFIQTQQLHA AMADTFLEHM CRLDIDSPPI TARNTGIICT
51  IGPASRSVET LKEMIKSGMN VARLNFSHGT HEYHAETIKN VRTATESFAS
101 DPILYRPVAV ALDTKGPEIR TGLIKGSGTA EVELKKGATL KITLDNAYME
151 KCDENILWLD YKNICKVVEV GSKIYVDDGL ISLQVKQKGA DFLVTEVENG
201 GSLGSKKGVN LPGAAVDLPA VSEKDIQDLK FGVEQDVDMV FASFIRKASD
251 VHEVRKVLGE KGKNIKIISK IENHEGVRRF DEILEASDGI MVARGDLGIE
301 IPAEKVFLAQ KMMIGRCNRA GKPVICATQM LESMIKKPRP TRAEGSDVAN
351 AVLDGADCIM LSGETAKGDY PLEAVRMQHL IAREAEAAIY HLQLFEELRR
401 LAPITSDPTE ATAVGAVEAS FKCCSGAIIV LTKSGRSAHQ VARYRPRAPI
451 IAVTRNPQTA RQAHLYRGIF PVLCKDPVQE AWAEDVDLRV NFAMNVGKAR
501 GFFKKGDVVI VLTGWRPGSG FTNTMRVVPV P
```

- **Identify « good peptides »**
  - **Good signal**
  - **Fragment well**
  - **Not modified**
- **Define « transition » :**

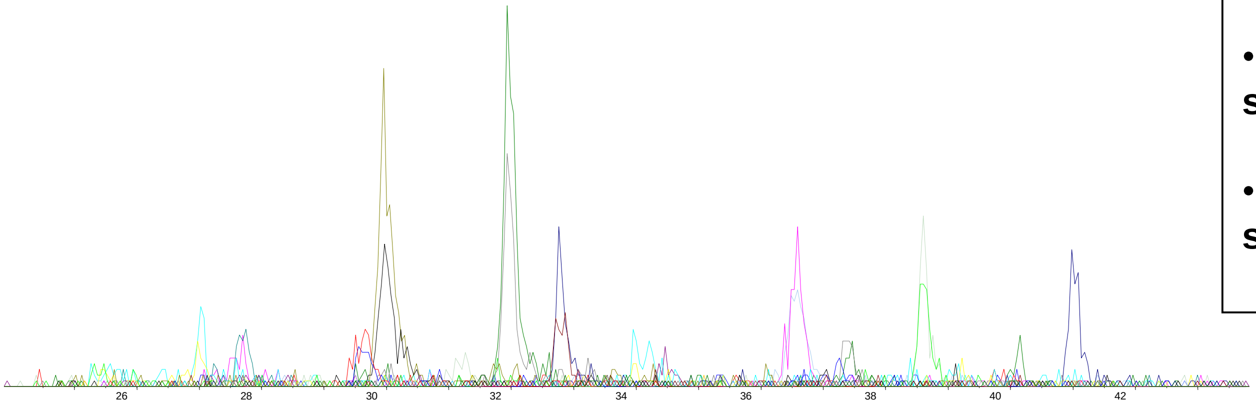
  m (precursor) / m (fragment)
- **Assemble list of transitions :**
  - **2-3 transitions / peptide**
  - **Min 2 peptides / protein**

Selected Reaction Monitoring on proteotypic peptides





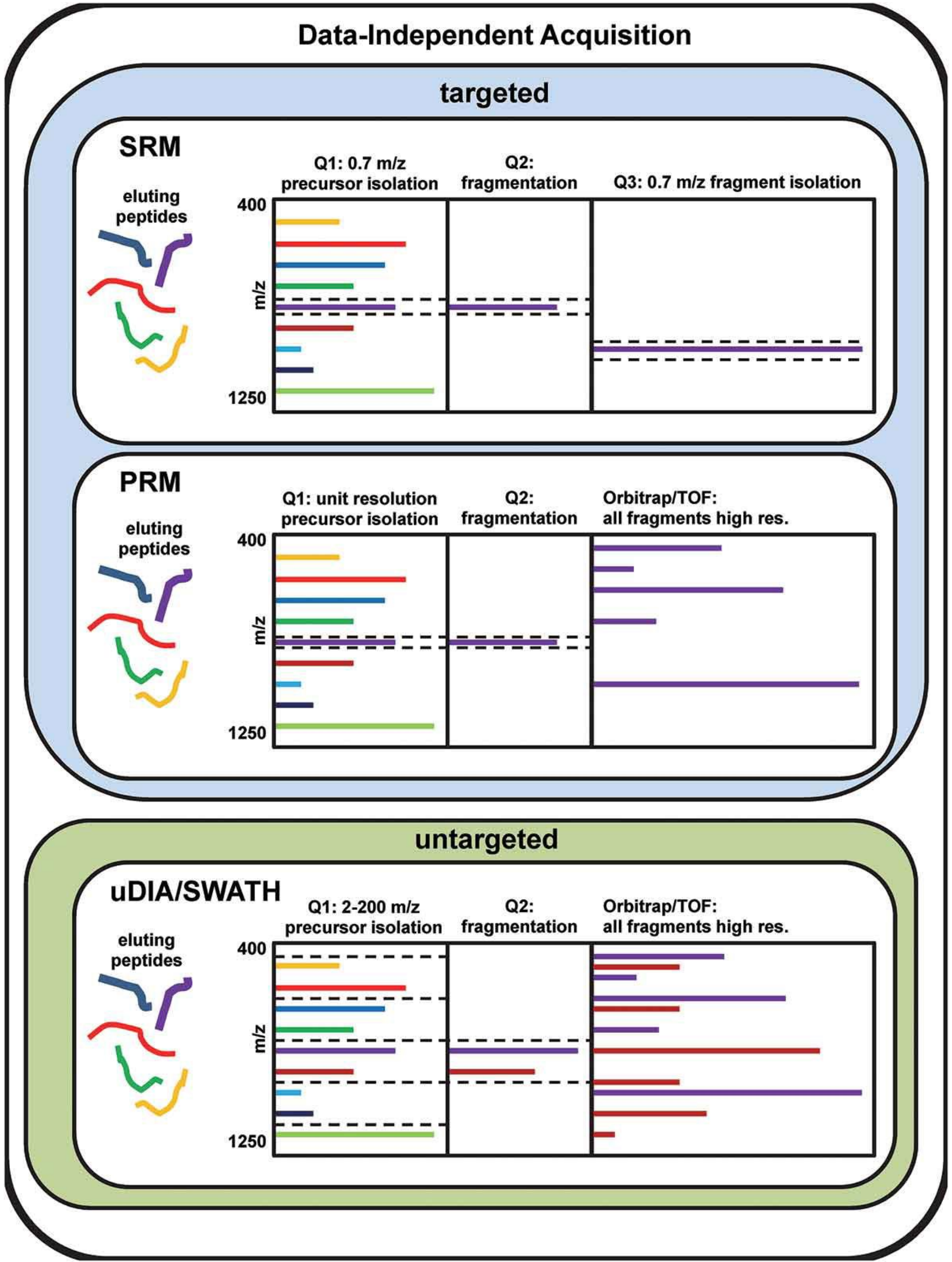
QQQ



- One trace / transition
- Transitions for same peptide should coelute
- Quantitation for peptides from same protein should correlate

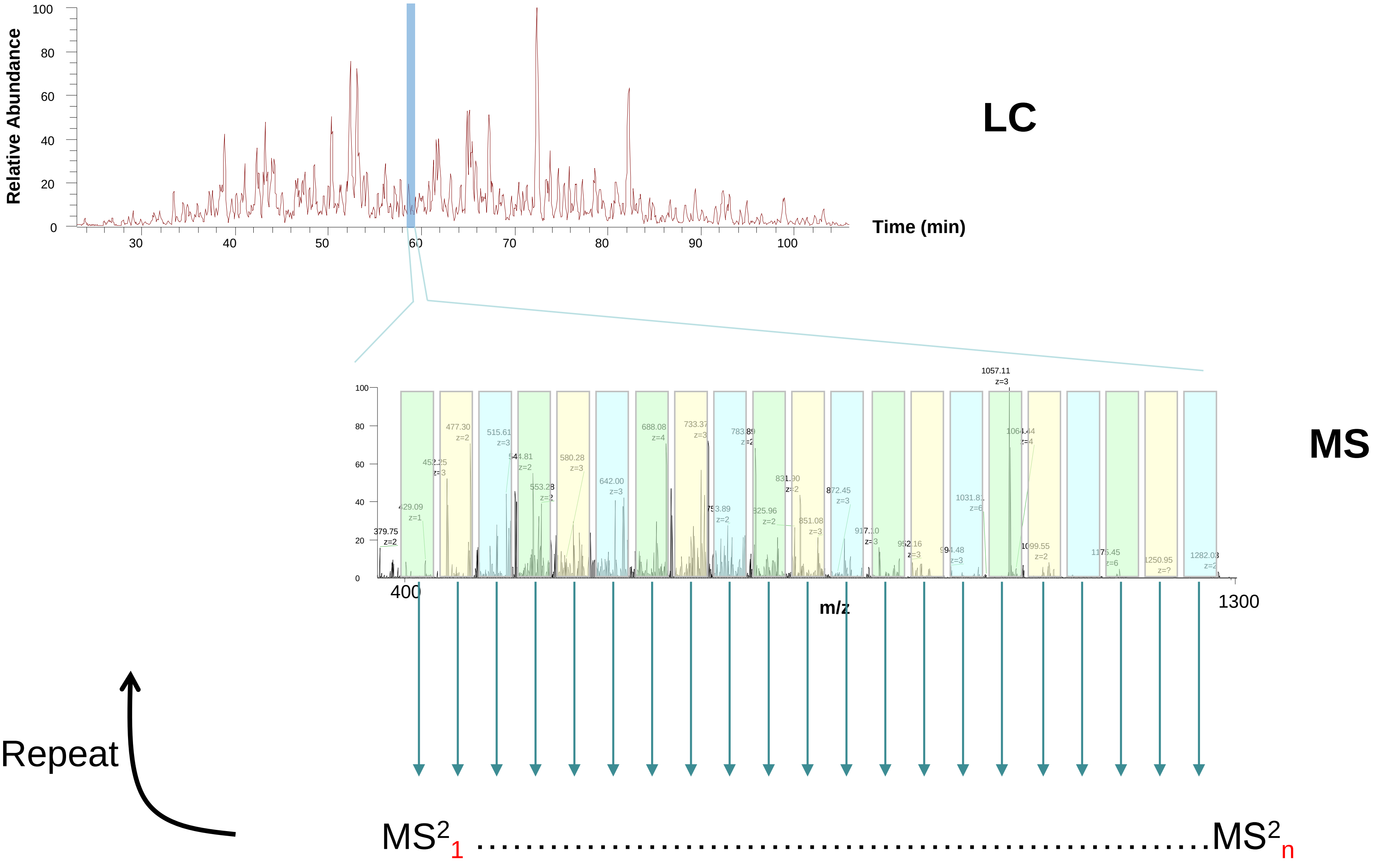
Targeted quantification and DIA

DIA: Data Independent Acquisition


Data-Independent Acquisition
targeted
SRM
eluting peptides
Q1: 0.7 m/z precursor isolation
Q2: fragmentation
Q3: 0.7 m/z fragment isolation
400
m/z
1250
PRM
eluting peptides
Q1: unit resolution precursor isolation
Q2: fragmentation
Orbitrap/TOF: all fragments high res.
400
m/z
1250
untargeted
uDIA/SWATH
eluting peptides
Q1: 2-200 m/z precursor isolation
Q2: fragmentation
Orbitrap/TOF: all fragments high res.
400
m/z
1250

Meyer JG, Schilling B.
Clinical applications of quantitative proteomics using targeted and untargeted data-independent acquisition techniques.
Expert Rev Proteomics. 2017 May;14(5):419-429.

DIA principle (simplest method) 1



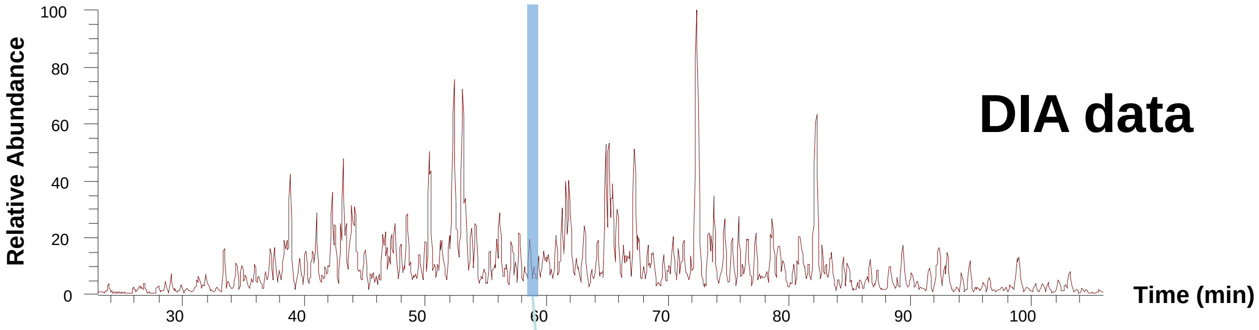
LC

MS

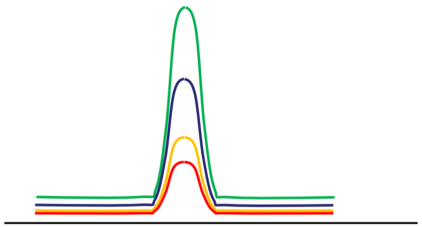
Repeat

MS^2_1 ....MS^2_n

DIA principle (simplest method) 2



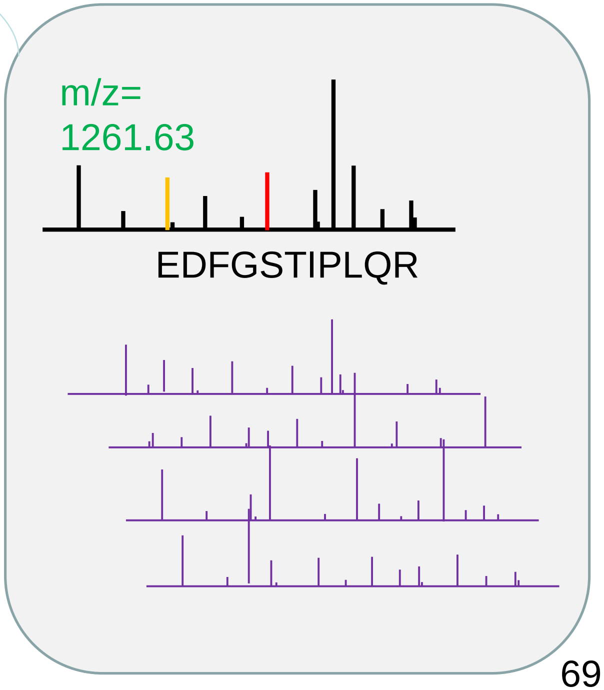
DIA data



Correlated XIC’s

Peptide «ID» + quant

Spectral library (DDA)



m/z=
1261.63

EDFGSTIPLQR

Data Dependent ⬄ Data Independent Acquisition

DDA

⊕

- Precursor isolation => specificity of MS/MS spectrum (?)
- Precursor isolation => max sensitivity (AGC)
- Flexible algorithm, «universal» method

⊖

- Semi-stochastic precursor selection => non reproducible
- Missed precursors => missing data

DIA

⊕

- Fragment ***everything*** => no missing data
- Sample record «complete», can be reinterrogated later
- More reproducible quantitation

⊖

- Large precursor windows => mixed MS/MS spectra
- Less sensitivity/specificity for weak precursors

Summary of key concepts

- **The Proteome : complexity, plasticity, dynamic range**
- **Proteomics : more challenging than genomics but direct access to cell functions**
- **LC & MS : many workflows to ID and quantify proteomes to depths of 5000 - 7000 proteins**

Take home message-1

- Many new possibilities in large scale protein analysis

PTM’s

- PTM are one of the most exciting and difficult « new » fields
- Huge variety and complexity of PTMs; no general workflow exists

Quantitative proteomics

- Quantitation is now feasible on a significant fraction of the proteome
- Several methods available; data quality and throughput are variable. Choice is often based on the experimental system and design

Take home message-2

- Some choices crucial for success:
    - Biological question : what are we looking for ?
    - Model system
    - Sample preparation (!)
        - Abundance of protein of interest
        - Complexity of mixture
        - Enrichment mechanism
    - Data analysis : *not soooooooooo easy* !
    - If we get results, can we interprete them ?
    - If we get results, are they going to be useful ?

Some good reviews

- Nesvizhskii AI, Vitek O, Aebersold R. (2007). Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat Methods, 4(10):787-97.*

  Review of proteomic analyses focused on statistical validation of data

- Jensen, O. (2006). Interpreting the protein language using proteomics. *Nat Rev Mol Cell Biol*, 7: 39-403.
- Witze, E.S., Old, W.M., Resing, K.A., & Ahn, N.G. (2007). Mapping protein post-translational modifications with mass spectrometry. *Nature Methods*, *4*(10): 798-806.
- Kim MS, Zhong J, Pandey A. (2016). Common errors in mass spectrometry-based analysis of post-translational modifications. Proteomics, 6(5):700-14.

  Give an overview of proteomics techniques used for PTM characterization in cells

- Bantscheff, M., Schirle, M., Sweetman, G., Rick, J. & Kuster, B. (2007). Quantitative mass spectrometry in proteomics: a critical review. *Anal Bioanal Chem, 389: 1017–1031.*
- Bantscheff, M., Lemeer, S., M., Savitski, MM. & Kuster, B. (2012). Quantitative mass spectrometry in proteomics: critical review update from 2007 to present. *Anal Bioanal Chem, 404: 939–965.*
- Eidhammer, I., Barsnes, H., Eide, GE., & Martens, L. (2013). Computational and Statistical Methods for Protein Quantification by Mass Spectrometry. *Wiley Ltd, 1st ed, Print ISBN: 9780470512975.*

  Review and comparison of analytical techniques used in quantitative proteomics

Contact

- www.unil.ch/paf

Activity of the facility, service fees, R&D, useful links, ….

- Manfredo.Quadroni@unil.ch
- Patrice.Waridel@unil.ch