Proteome quantitation by Mass Spectrometry (MS)

A short guide to data processing and interpretation

Protein Analysis Facility
University of Lausanne
Faculty of Biology and Medicine
Lausanne, Switzerland

www.unil.ch/paf/

Version 4 - April 12th, 2023

Goals of this guide

- Explain last part of data analysis pipeline, with special attention to processing steps that impact final results

- Describe format of data sent to users

- Describe options for downstream data processing and interpretation

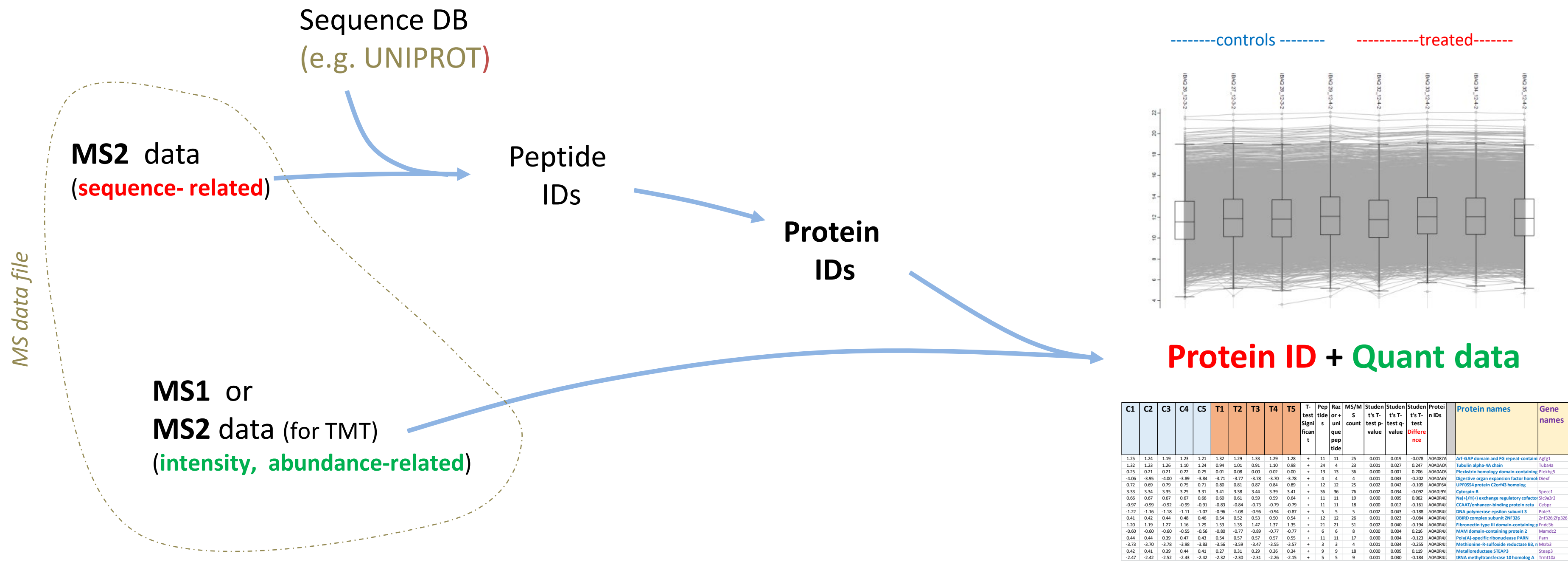
- Answer commonly asked questions

- Note : a separate guide will be available for PTM-specific datasets (e.g. phosphoproteomics)

Index

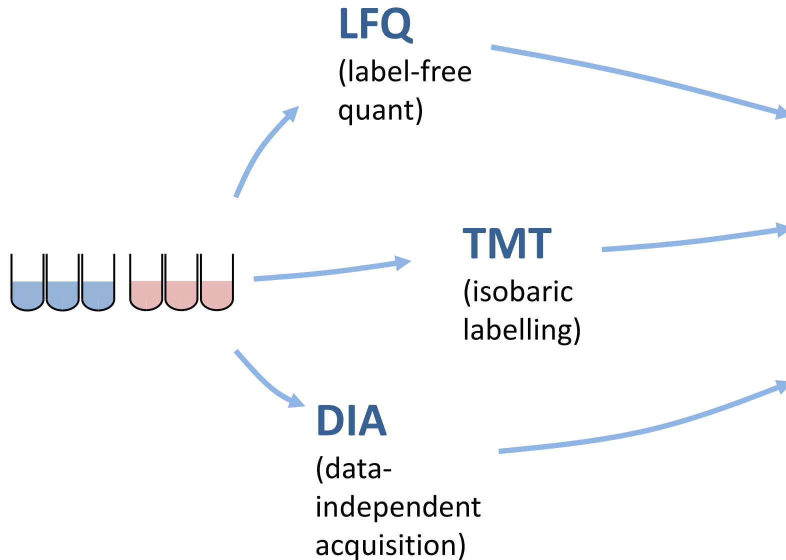
- Description of general MS workflow … 4
- Overview of data processing steps …6
- Description of processed output (main table) …12
- General facts & things to remember …15
- What if things do not work out …17
- Annotation enrichment : concepts …18
- Annotation enrichment : output, usage of tables …19
- Publication & data submission to repositories …23
- Misc info …24
- Appendix : MaxQuant variables explained …25

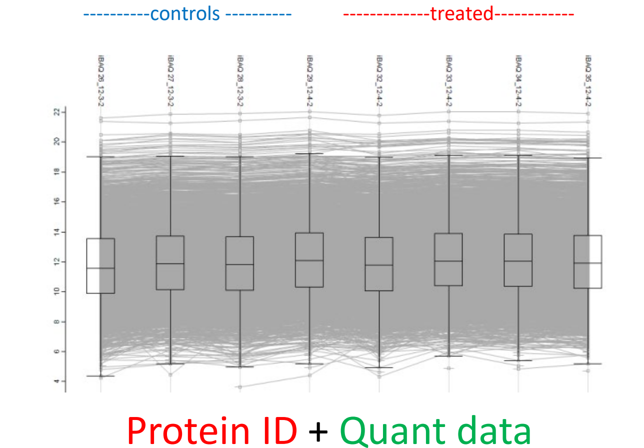
Nuts and bolts : Mass Spec data and measurables



MS analysis simultaneously produces different types of information (MS1 and MS2), used for different purposes

Different MS pipelines, same type of output

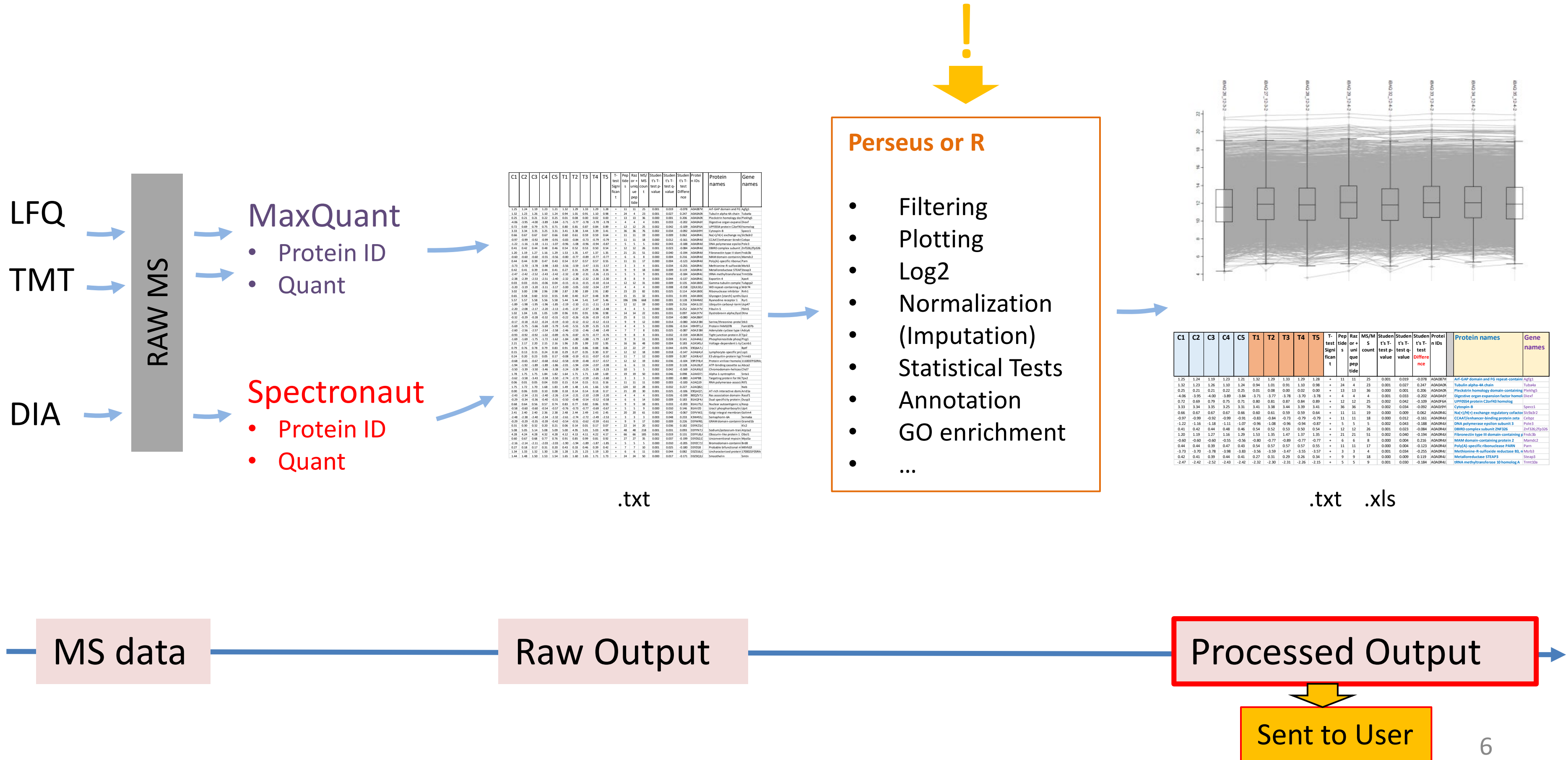




C1	C2	C3	C4	C5	T1	T2	T3	T4	T5	T-test Significant	Peptides	Razor + unique peptide	MS/MS count	Student's T-test p-value	Student's T-test q-value	Student's T-test Difference	Protein IDs		Protein names	Gene names
1.25	1.24	1.19	1.23	1.21	1.32	1.29	1.33	1.29	1.28	+	11	11	25	0.001	0.019	-0.078	A0A087W		Arf-GAP domain and FG repeat-containi	Agfg1
1.32	1.23	1.26	1.10	1.24	0.94	1.01	0.91	1.10	0.98	+	24	4	23	0.001	0.027	0.247	A0A0A0N		Tubulin alpha-4A chain	Tuba4a
0.25	0.21	0.21	0.22	0.25	0.01	0.08	0.00	0.02	0.00	+	13	13	36	0.000	0.001	0.206	A0A0A0N		Pleckstrin homology domain-containing	Plekhg5
-4.06	-3.95	-4.00	-3.89	-3.84	-3.71	-3.77	-3.78	-3.70	-3.78	+	4	4	4	0.001	0.033	-0.202	A0A0A6Y		Digestive organ expansion factor homol	Diexf
0.72	0.69	0.79	0.75	0.71	0.80	0.81	0.87	0.84	0.89	+	12	12	25	0.002	0.042	-0.109	A0A0F6A		UPF0554 protein C2orf43 homolog	
3.33	3.34	3.35	3.25	3.31	3.41	3.38	3.44	3.39	3.41	+	36	36	76	0.002	0.034	-0.092	A0A0J9YU		Cytospin-B	Specc1
0.66	0.67	0.67	0.67	0.66	0.60	0.61	0.59	0.59	0.64	+	11	11	19	0.000	0.009	0.062	A0A0R4I2		Na(+)/H(+) exchange regulatory cofactor	Slc9a3r2
-0.97	-0.99	-0.92	-0.99	-0.91	-0.83	-0.84	-0.73	-0.79	-0.79	+	11	11	18	0.000	0.012	-0.161	A0A0R4J(		CCAAT/enhancer-binding protein zeta	Cebpz
-1.22	-1.16	-1.18	-1.11	-1.07	-0.96	-1.08	-0.96	-0.94	-0.87	+	5	5	5	0.002	0.043	-0.188	A0A0R4J(		DNA polymerase epsilon subunit 3	Pole3
0.41	0.42	0.44	0.48	0.46	0.54	0.52	0.53	0.50	0.54	+	12	12	26	0.001	0.023	-0.084	A0A0R4J(		DBIRD complex subunit ZNF326	Znf326;Zfp326
1.20	1.19	1.27	1.16	1.29	1.53	1.35	1.47	1.37	1.35	+	21	21	51	0.002	0.040	-0.194	A0A0R4J(		Fibronectin type III domain-containing p	Fndc3b
-0.60	-0.60	-0.60	-0.55	-0.56	-0.80	-0.77	-0.89	-0.77	-0.77	+	6	6	8	0.000	0.004	0.216	A0A0R4J(		MAM domain-containing protein 2	Mamdc2
0.44	0.44	0.39	0.47	0.43	0.54	0.57	0.57	0.57	0.55	+	11	11	17	0.000	0.004	-0.123	A0A0R4J(		Poly(A)-specific ribonuclease PARN	Parn
-3.73	-3.70	-3.78	-3.98	-3.83	-3.56	-3.59	-3.47	-3.55	-3.57	+	3	3	4	0.001	0.034	-0.255	A0A0R4J1		Methionine-R-sulfoxide reductase B3, m	Msrb3
0.42	0.41	0.39	0.44	0.41	0.27	0.31	0.29	0.26	0.34	+	9	9	18	0.000	0.009	0.119	A0A0R4J1		Metalloreductase STEAP3	Steap3
-2.47	-2.42	-2.52	-2.43	-2.42	-2.32	-2.30	-2.31	-2.26	-2.15	+	5	5	9	0.001	0.030	-0.184	A0A0R4J2		tRNA methyltransferase 10 homolog A	Trmt10a

The final format of the data table is often similar, independently of the pipeline adopted

Data analysis steps and software tools



From Raw output to Processed output - 1

Filtering :

- remove *reverse* hits
- remove contaminants
- remove proteins with low data quality

In large datasets, a search is performed against a «decoy» database made of *reverse* sequences, to determine the frequence of false positive ID's. Scoring parameters are then adjusted to have 1% FDR in the final list.

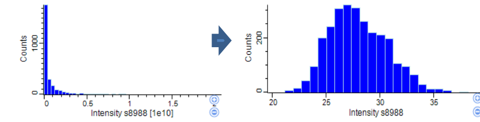
E.g. Keratins, Trypsin, Bovine Serum Albumin,...

For ex. proteins identified by only 1 peptide in only 1 sample

Transform values :

- Log2 transform quant values

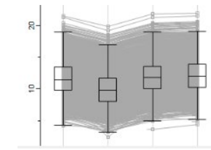
Obtain a normal distribution, make plotting easier



Plotting : look at

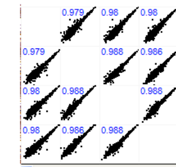
- medians , distributions
- negative and positive control proteins
- correlation between samples, clustering

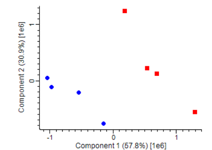
Detect loading differences, measurement biases



Assess sensitivity, biological QC

QC of conditions. Find outliers. Any groups emerging ?





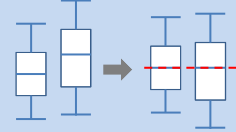
From Raw output to Processed output - 2

Normalization :

Remember : it is usually assumed that most of the proteome does not change (stable background)

- Correct for differences in total loading
- normalize based on intensity of a protein of interest

Simplest method : equalize median of populations

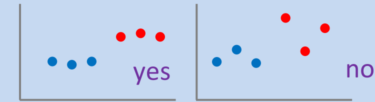


In case of protein complexes, organelles etc

Statistical tests : find significant differences

- Define groups/conditions
- Student's T-test or Welch Test
- Large datasets ➔ multiple testing correction is necessary

Difference between groups must be larger than variability within groups.



Statistical power (# of replicates) is crucial.

Correction of p-value => q-value.
Benjamini-Hochberg or other methods

From Raw output to Processed output - 3

Add functional annotation :

annotation is useful but also redundant and not always perfect

- Add Gene Ontology (GO) , KEGG, Corum, Reactome,…

Annotation enrichment analysis :

A sensitive method that can detect broad changes of small amplitude

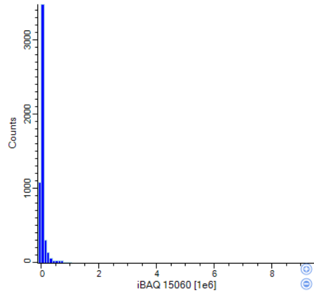
- Are the proteins that change enriched in some pathways, localization, functions ?

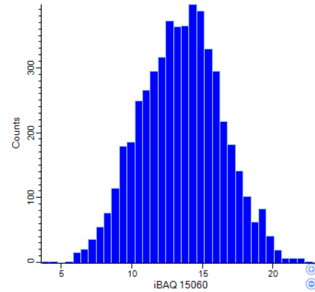
Many terms, long lists…risk of false positives ! Multiple testing correction is needed here, too.

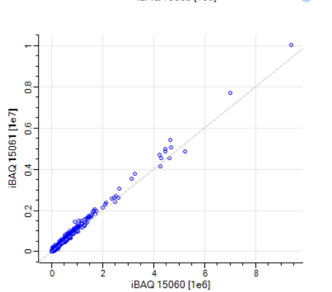
…but why log intensity data ?

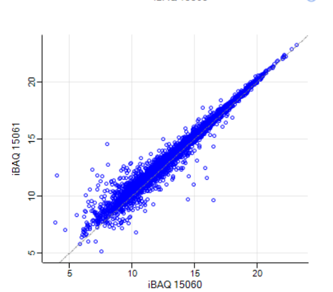
Raw (LINEAR) data :
- Vast dynamic range
- Asymmetric distributions

Transformed (log) data :
- Manageable dyn. range (plots)
- More symmetric distributions









Reminder : log-transformed intensities :

Multiplication (LIN) ➔ sum (log)
Division (LIN) ➔ subtraction (log)

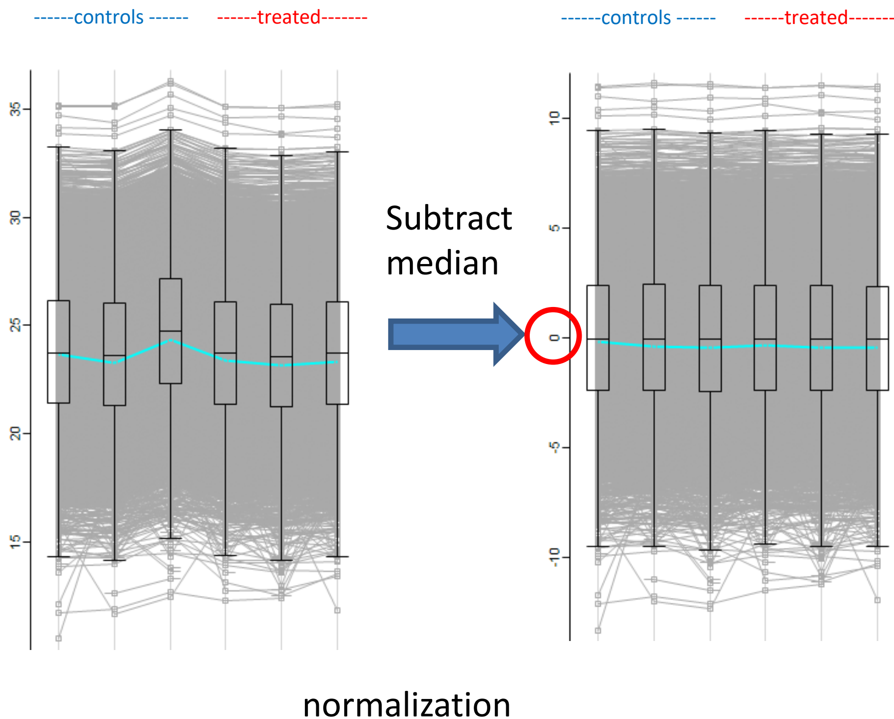
i.e.

LIN : $A / B = X$ [FOLD change]

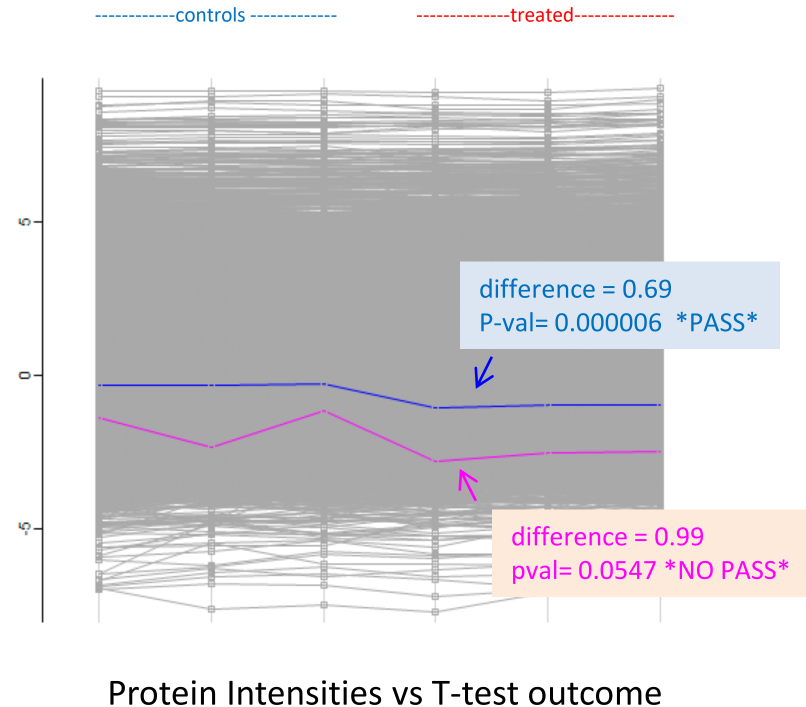
$\log_2$: $\log_2(A) - \log_2(B) = \log_2(X)$ [$\log_2$ diff]

$\log_2$ diff = 1.0 ➔ A/B =2
$\log_2$ diff = -1.0 ➔ A/B =0.5

Examples of profile plots for visualization



normalization



Protein Intensities vs T-test outcome

Content of Main Result Tables

The exact identity and naming of columns can vary

Quant values for conditions. Depending on pipeline, these can be «Intensity» or «LFQ» or «iBAQ» or «PG:Quantity»

Is protein group passing Statistical test ? Usually with Q-val<0.05

Number of peptides used for ID or quant. Includes shared peptides

Unique peptides for this protein group

-log(10) of T-test p-value (uncorrected)

Q val = p-value after multiple testing correction

Difference of means of conditions. Equivalent to fold change, but in LOG2 scale

Each line is a PROTEIN GROUP, including one or more database sequences, in case they cannot be discriminated by the set of assigned peptides

LFQ - Ctr R1	LFQ - Ctr R4	LFQ - Ctr R3	LFQ KO - R1	LFQ KO - R2	LFQ KO - R3	Student's T-test Significant Ctr_KO	GOBP name	GOMF name	GOCC name	KEGG name	Peptides	Razor + unique peptides	Unique peptides	Sequence coverage [%]	Unique + razor sequence coverage [%]	Unique sequence coverage [%]	Mol. weight [kDa]	Q-value	Score	Intensity	MS/MS count	iBAQ	Razor + unique peptides 11905	Razor + unique peptides 11906	Razor + unique peptides 11907	Razor + unique peptides 11908	Razor + unique peptides 11909	Razor + unique peptides 11910	Razor + unique peptides 11911	Razor + unique peptides 11912	Razor + unique peptides 11913	Razor + unique peptides 11914	Razor + unique peptides 11915	Razor + unique peptides 11916	-Log Student's T-test p-value Ctr_ KO	Student's T-test q-value Ctr_KO	Student's T-test Difference Ctr_KO	Protein IDs	Majority protein IDs	Protein names	Gene names
-0.28	-0.29	-0.25	-1.00	-0.95	-0.95	+	adipose	5'-deoxy	cell part;chromat		**5**	**5**	**5**	46.8	46.8	46.8	11.832	0	33.63	31.44	**80**	29.1	5	5	5	5	5	5	3	3	3	3	3	3	5.161	0.035	0.694	P52926;F5	P52926;F5	High mobility group protein HMGI-	HMGA2
-2.32	-2.31	-2.30	-2.03	-2.04	-2.01	+	actin cyt	enzyme	cell junction;cell		**22**	**22**	**22**	18.7	18.7	18.7	200.03	0	61.83	31.41	**107**	24.9	6	5	6	5	5	6	17	18	18	18	19	17	4.810	0.039	-0.285	O43166;F5	O43166;F5	Signal-induced proliferation-associ	SIPA1L1
0.37	0.39	0.37	0.59	0.59	0.61	+	anatomi	activin r	cell junc	Tight jur	**9**	**9**	**8**	33.5	33.5	31.9	54.461	0	105.5	32.4	**62**	28.3	9	9	9	8	9	9	4	4	4	4	4	3	4.539	0.049	-0.221	P46937;HC	P46937;HC	Transcriptional coactivator YAP1	YAP1
-1.97	-1.95	-1.96	-1.68	-1.63	-1.65		biosynth	catalytic	cell part;cytoplas		**7**	**7**	**7**	26.7	26.7	26.7	38.155	0	21.91	30.53	**36**	26.6	3	3	3	3	3	3	5	4	5	3	6	6	4.260	0.070	-0.306	Q5T440	Q5T440	Putative transferase CAF17, mitoch	IBA57
5.15	5.15	5.15	4.96	5.00	4.97		cellular	binding;	cell part	Systemi	**33**	**33**	**33**	58.6	58.6	58.6	46.836	0	308.2	37.49	**730**	33	32	32	32	32	32	32	31	29	29	31	30	29	4.087	0.083	0.180	P05455;E7	P05455	Lupus La protein	SSB
-0.57	-0.55	-0.56	-0.37	-0.32	-0.34		catabolic process		cell part	DNA rep	**5**	**5**	**5**	36.8	36.8	36.8	26.302	0	10.34	31.79	**29**	28.2	5	4	4	4	4	4	2	2	2	2	2	2	3.902	0.102	-0.214	E9PN81;Q	E9PN81;Q	Ribonuclease H2 subunit C	RNASEH2C
1.57	1.57	1.59	1.67	1.68	1.68		biologic	catalytic	cell part;cytoplas		**20**	**20**	**20**	35	35	35	77.528	0	112.2	33.85	**214**	28.5	16	15	14	15	13	14	17	15	16	16	17	16	3.809	0.102	-0.098	Q6ZRP7;H	Q6ZRP7;H	Sulfhydryl oxidase 2;Sulfhydryl oxi	QSOX2
0.10	0.08	0.11	-0.22	-0.21	-0.28		carboxyl	adenyl n	cell part;cytoplas		**13**	**13**	**13**	35.8	35.8	35.8	63.652	0	83.22	32.43	**75**	27.8	11	10	11	9	10	9	8	8	9	8	8	8	3.767	0.102	0.333	E9PF16;Q	E9PF16;Q	Acyl-CoA synthetase family membe	ACSF2
-2.54	-2.70	-2.53	-1.84	-1.76	-1.78		anatomi	binding;	apical pa	Basal cel	**5**	**5**	**4**	25.3	25.3	21.2	39.917	0	27.07	30.3	**14**	26	4	3	4	3	4	3	3	3	3	3	2	3	3.706	0.102	-0.795	I3L2W9;O	I3L2W9;O	Segment polarity protein dishevell	DVL2
2.18	2.23	2.23	2.00	1.97	1.97		axon gui	binding;	adheren	Cell adh	**23**	**23**	**23**	47.5	47.5	47.5	65.102	0	122.2	34.87	**266**	30	19	19	20	19	20	19	21	21	22	21	20	21	3.631	0.102	0.238	Q13740;F5	Q13740;F5	CD166 antigen	ALCAM

LOG2 scale ! Normalized values can be negative (=lower than median)

Functional annotation (many terms per line)

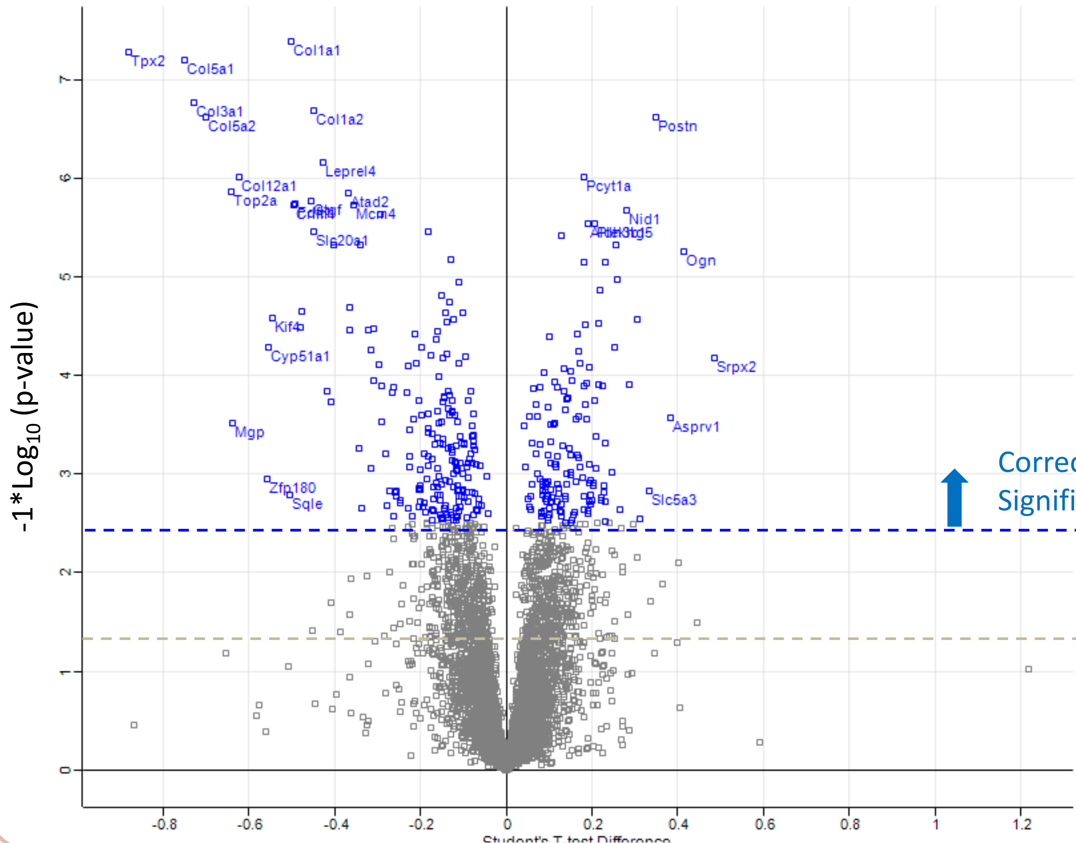
Includes shared peptides but assigned only to highest scoring protein group. These are the peptides used for quantitation. Use this for quality filtering

MS/MS count : Can also be used for filtering/validation if column is present

Database identifier. First one is the best.

Volcano plot

A commonly used way to represent statistical significance (*y-axis*) and fold change (*x-axis*) in a single plot



-1*Log$_{10}$ (p-value)

Corrected p-val (q-value) <0.05
Significant proteins

p-val <0.05
Not sufficiently stringent

Student's T-test difference (Log$_2$ scale)

Missing data (MD) and imputation

- Quantitative proteomics data always contain missing data points. These can result from i) stochasticity of data acquisition in MS ii) peptides/proteins that are near the limit of detection and below in some samples iii) truly absent proteins. Unfortunately it is impossible to discriminate between these events.

- The percentage of MD is variable, some types of data contain more MD (LFQ, iBAQ) , others less (TMT, DIA)

- MD's are a data analysis problem : fold changes cannot be calculated, proteins do not appear on plots.

- Data imputation is the substitution of MD with values used for computations. Imputed values are often chosen at the low end of the distribution, assuming that the reason for the MD is very low signal intensity.

- Imputation has to be applied with caution and on a limited scale. Imputed data should be flagged to be easily identified (and removed) later.

General facts and things to remember

- Proteomes tend to change less than transcriptomes, i.e. the amplitudes of changes are often smaller and the changes slower.

- Reliability of quantitation is lower for weak signals and for proteins with a low number of identified/quantified peptides. These are mostly low abundance proteins.

- Statistical power is crucial to find significantly different proteins. Power is mostly determined by the number of biological replicates. $n=3$ is the minimum for having reliable (*and publishable*!) data. $n>3$ gives much better statistics, especially when quantifying mild changes or when variability is higher, e.g. $n>3$ is recommended for studies on individual animals (mice).

Evaluation of processed output

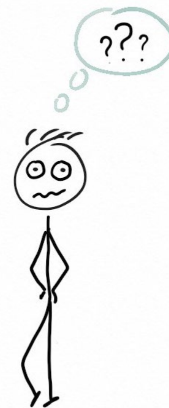
Defining a list of candidate differentially expressed proteins :

- Protein candidates that pass statistical test should be evaluated critically using quality parameters, e.g. number of quantitated peptides or ratios
- Also consider filtering by minimal fold change (0.5x or more)
- Manual check of MS data for validation is possible but very time consuming
- <u>Repetition of experiments is the best confirmation.</u>

Caveat : # peptides vs. # unique peptides

Caveat : always remember if values are in linear or log scale

…ehm, it did not work…



- Sometimes no statistically significant proteins emerge. Most frequent reasons :
  - only small differences between conditions
  - insufficient statistical power (*n* too low)
  - high variability within groups.
    - ➔ Often more than one of these factors combined.

- How to move on ?
  - It does not necessarily mean there is no difference, just the experiment was not perfect.
  - Select best candidates (best fold change & p-val, >3 peptides) and try to quantify them in a targeted manner (WB, FACS, ELISA,MS)
  - For global trends : look at annotation enrichment analysis
  - Repeat, reducing variability or increasing *n*

Annotation enrichment – basic concepts

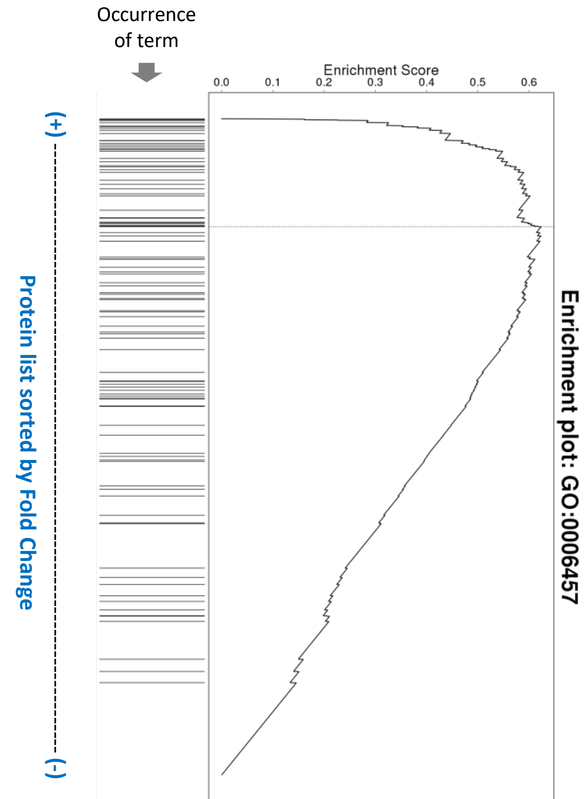
Principle (for example GSEA algorithm) :

1. Annotation terms from a controlled vocabulary (for ex. GO terms, http://geneontology.org/) are added to each protein in the list. These terms describe functional or structural characteristics, interactions, pathways or other features.

2. The list is first sorted by a relevant parameter resulting from the experiment. This is typically fold change between conditions

3. An algorithm is used to determine for each annotation term (for ex. «Ribosomal protein») if it is significantly enriched towards the top or the bottom of the list, e.g. tends to be more associated with proteins that increase or decrease between conditions.

4. A p-value is calculated and should be corrected for multiple testing in case of large datasets.



Note : the algorithm used by Perseus is not GSEA but is conceptually similar
Ref : Cox, J., & Mann, M. (2012) *BMC Bioinformatics*, *13 Suppl 1*(Suppl 16), S12. https://doi.org/10.1186/1471-2105-13-S16-S12

Annotation enrichment – results table

Parameter used for sorting the table. Usually «difference» used (~fold change).

Number of protein groups in the list having this annotation term

P-value for enrichment of this term

Median of «difference» for each GO term. In Log2 as usual

Column	Type	Name	Size	Score	P value	Benj. Hoch. FDR	Mean	Median (>0 means decreasing in KO and viceversa)
Student's T-test Difference Ctr_KO	GOCC name	**mitochondrial outer membrane translo**	7	**0.769**	0.000433	0.009	0.281	0.27
Student's T-test Difference Ctr_KO	GOCC name	**proton-transporting ATP synthase comp**	8	**0.681**	0.000854	0.015	0.182	0.13
Student's T-test Difference Ctr_KO	GOMF name	**SNAP receptor activity**	20	**0.517**	0.000065	0.008	0.143	0.13
Student's T-test Difference Ctr_KO	GOBP name	**protein transmembrane transport**	14	**0.563**	0.000270	0.018	0.143	0.12
Student's T-test Difference Ctr_KO	GOCC name	**mitochondrial proton-transporting ATP**	13	**0.693**	0.000015	0.001	0.161	0.12
Student's T-test Difference Ctr_KO	GOCC name	**proton-transporting ATP synthase comp**	13	**0.693**	0.000015	0.001	0.161	0.12
Student's T-test Difference Ctr_KO	GOMF name	**secondary active transmembrar e transp**	26	**0.537**	0.000002	0.000	0.109	0.11
Student's T-test Difference Ctr_KO	GOMF name	**symporter activity**	16	**0.543**	0.000174	0.018	0.108	0.11
Student's T-test Difference Ctr_KO	GOBP name	**response to pH**	13	**0.699**	0.000013	0.002	0.300	0.11
Student's T-test Difference Ctr_KO	GOCC name	**mitochondrial small ribos ubunit**	19	**0.533**	0.000059	0.002	0.075	0.11

GOBP = biological process
GOCC=cellular compartment
GOMF=molecular function

Terms are often redundant, also due to hierarchical organization of GO terms

Corrected P-val (Benjamini-Hochberg method). A cutoff of 0.02-0.05 on this value is applied to generate the list

Useful Trick : Retrieving sets of proteins with a given GO annotation

(from a .txt or .xls table)

Column	Type	Name	Size	Score	P value	Benj. Hoch. FDR	Mean	Median (>0 means decreasing in KO and viceversa)
Student's T-test Difference Ctr_KO	GOCC name	**mitochondrial outer membrane translo**	7	**0.769**	0.000433	0.009	0.281	0.27
Student's T-test Difference Ctr_KO	GOCC name	**proton-transporting ATP synthase comp**	8	**0.681**	0.000854	0.015	0.182	0.13
Student's T-test Difference Ctr_KO	GOMF name	**SNAP receptor activity**	20	**0.517**	0.000065	0.008	0.143	0.13
Student's T-test Difference Ctr_KO	GOBP name	**protein transmembrane transport**	14	**0.563**	0.000270	0.018	0.143	0.12
Student's T-test Difference Ctr_KO	GOCC name	**mitochondrial proton-transporting ATP**	13	**0.693**	0.000015	0.001	0.161	0.12
Student's T-test Difference Ctr_KO	GOCC name	**proton-transporting ATP synthase comp**	13	**0.693**	0.000015	0.001	0.161	0.12
Student's T-test Difference Ctr_KO	GOMF name	**secondary active transmembrane trans**	26	**0.537**	0.000002	0.000	0.109	0.11
Student's T-test Difference Ctr_KO	GOMF name	**symporter activity**	16	**0.543**	0.000174	0.018	0.108	0.11
Student's T-test Difference Ctr_KO	GOBP name	**response to pH**	13	**0.699**	0.000013	0.002	0.300	0.11
Student's T-test Difference Ctr_KO	GOCC name	**mitochondrial small ribosomal subunit**	19	**0.533**	0.000059	0.002	0.075	0.11

How to get these 13 proteins ??

Retrieving sets of proteins with a given GO annotation -2

1. Go to excel table
2. Select column with annotation of interest (ex GOBP)
3. Go to «data» and click «Filter»

Phospho(Y) sites Schairer 9881-82-89-90-96-97_filter_GO_M2.x

FILE | HOME | Menu | INSERT | PAGE LAYOUT | FORMULAS | DATA | REVIEW | VIEW | ADD-INS

Get External Data: From Access | From Web | From Text | From Other Sources | Existing Connections

Connections: Refresh All | Connections | Properties | Edit Links

Sort & Filter: Sort | Filter | Clear | Reapply | Advanced

Data Tools: Text to Columns | Flash Fill | Remove Duplicates | Data Validation | Consolida

F1 | GOBP name

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Ratio H/L normalized 9889___	Ratio H/L normalized 9890___	Amino acid	Charge	T-test Significant p005	GOBP name	GOMF name	GOCC name	id	Localization prob	PEP	Score	Delta score	Score for localization	Mass error [ppm]	Intensi
56	0.498	0.693	Y	2		biological	regulation	cell part;c	45	1	2.88E-08	121.83	109.25	110.14	0.7665	35117
57	0.401	0.750	Y	2		apoptosis	enzyme ac	cell juncti	317	1	0.02792	43.083	15.558	43.083	0.154	18009
58	0.190	0.954	Y	3		biosynthe	binding;ca	cell part;c	424	0.99989	3.97E-07	71.692	67.747	71.692	0.56951	16171
59	0.225	0.915	Y	3		assembly	binding;rib	Cajal body	383	0.99518	0.00031	72.652	72.652	72.652	0.09986	26249
60	0.313	0.824	Y	2		alcohol m	catalytic a	brush bord	354	1	0.013	42.001	42.001	42.001	1.1592	15493
61	-0.683	1.809	Y	4		axon guid	binding;ca	adherens_	474	1	2.73E-15	87.659	87.58	70.315	1.0117	64458
62	-0.013	1.113	Y	2		biological	binding;ca	adherens_	94	0.99971	0.00587	61.999	18.753	61.999	0.27139	19941
63	0.329	0.752	Y	2		apoptosis	enzyme ac	cell juncti	319	1	5.22E-09	92.063	85.024	87.363	0.65438	34889
64	-0.121	1.191	Y	3		cell cycle	cytokinesis	cell divisio	332	1	4.97E-08	87.18	87.18	76.588	0.29336	39266
65	-0.210	1.213	Y	3					338	0.99245	7.02E-06	49.089	49.089	49.089	1.0696	17460
66	-1.370	2.332	Y	2		antigen pr	binding;ca	cell part;c	305	1	0.00111	86.794	72.637	86.794	0.17635	19440
67	-0.079	1.029	Y	3		biosynthe	binding;ca	cell part;c	423	1	7.16E-05	62.924	56.761	62.924	0.11816	16059
68	-0.143	1.084	Y	3		actin cyto	adenyl nu	cell juncti	109	1	2.39E-20	94.384	90.986	62.663	0.75242	1.4E+
69	-0.131	1.043	Y	3			receptor a	cell part;e	397	1	2.53E-21	102.6	102.6	99.366	0.70128	1.1E+
70	-0.043	0.947	Y	2		alcohol bi	binding;ca	cell part;c	112	1	3.17E-05	96.331	91.88	89.247	-0.0605	1.4E+
71	NaN	0.427	Y	3		alcohol m	adenyl nu	cell part;c	127	1	2.72E-07	66.965	66.965	66.965	0.93665	11423
72	0.175	0.664	Y	2		actin cyto	actin bind	adherens_	78	1	3.60E-07	82.417	68.647	82.417	0.78895	15838
73	-2.307	1.738	Y	2		activation	adenyl nu	cell juncti	196	1	4.87E-07	78.908	78.908	78.908	0.8931	20781
74	0.265	0.522	Y	3		amine me	binding;ca	cell part;c	97	0.99687	6.57E-11	86.455	86.455	86.455	1.2113	17814
75	0.435	0.331	Y	4			receptor a	cell part;e	396	1	1.03E-14	119.45	110.58	66.845	0.69284	4.3E+
76	NaN	0.369	Y	2		actin cyto	actin bind	actin cyto:	321	1	0.01028	73.985	65.453	73.985	0.15248	11070
77	-0.293	1.030	Y	2		axon guid	catalytic a	cell part;e	138	1	5.62E-17	109.6	109.6	109.6	0.5939	9.8E+
78	-0.509	1.245	Y	2		amine tra	active tran	cell part;c	360	1	3.10E-05	106.42	93.896	106.42	0.23995	2.2E+
79	-0.432	1.148	Y	3		anatomica	binding;ca	adherens_	341	1	5.46E-05	54.501	49.952	54.501	-0.01162	13500
80	0.042	0.656	Y	3				cell part;c	442	1	2.01E-08	72.421	72.421	72.421	0.63651	76487
81	-0.271	0.963	Y	2		actin cyto	actin bind	actin cyto:	422	1	2.74E-05	85.958	83.914	85.958	1.328	1.2E+

Phospho(Y) sites Schairer 9881-

Retrieving sets of proteins with a given GO annotation -3

1. Go to pull-down menu

2. Select «text filters» and «contains»

3. Fill in the «contains» field the name of the annotation you are looking for

4. Click OK

Publication and data submission to repositories

- Please **read** our guidelines on acknowledgements of facility work and authorship :
  https://www.unil.ch/paf/home/menuguid/policy.html

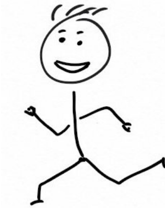
- Remember to ask us for Math & Meth descriptions, rather sooner than later. Three years afterwards it may prove difficult to recover all the information relative to your experiment.

- Upload of raw MS data and processed data to repositories has become highly recommended or compulsory for most journals. As a repository we recommend proteomexchange.org (part of PRIDE https://www.ebi.ac.uk/pride/).

- Submission to the data repository requires i) raw MS data ii) project title, description, metadata iii) Materials and Methods iv) protein ID and quant

- Currently (as of March 2023) we will still do the data preparation and upload for our users but we need sufficient notice, i.e. one week minimum before manuscript submission. In case of really large datasets we may need more time and we may have to charge you for some of the work involved.

Misc info and references

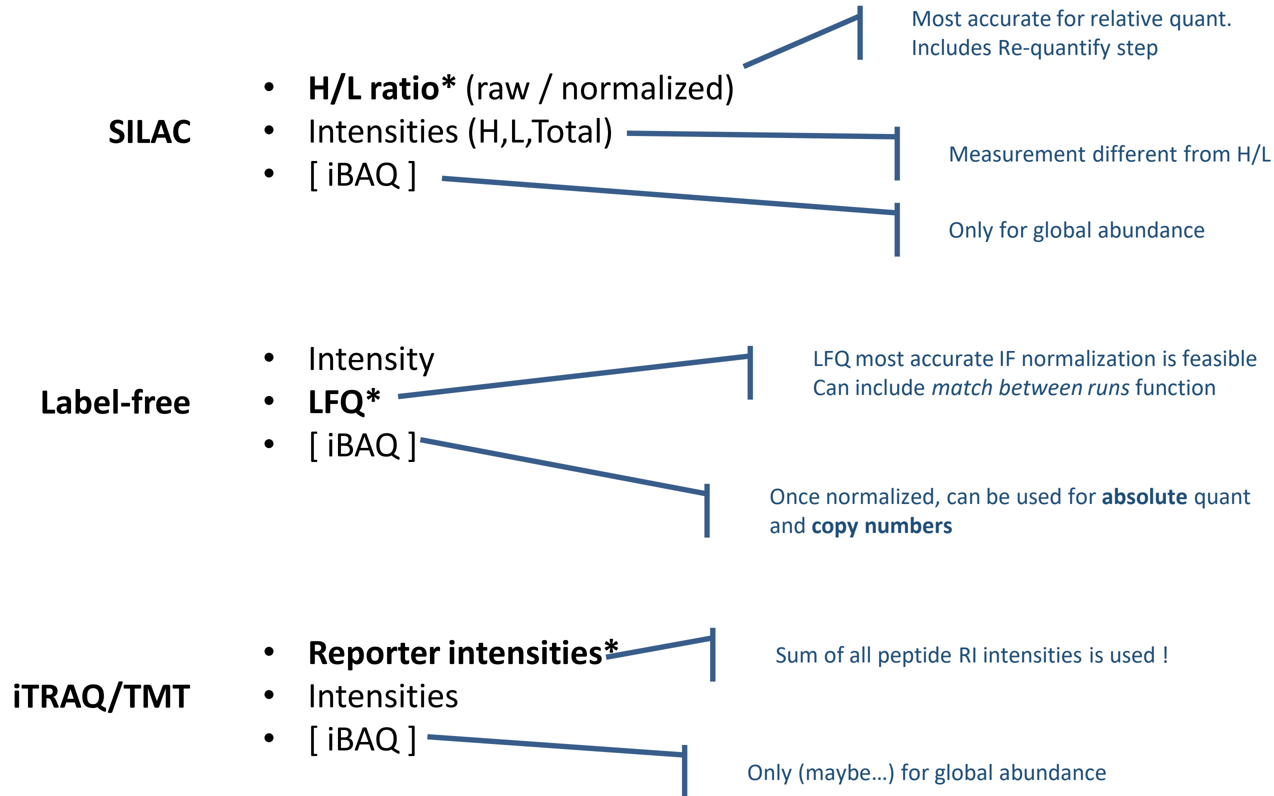


- MaxQuant (https://maxquant.net/maxquant/)
- Spectronaut (https://biognosys.com/technology/publications)
- **Perseus** is a free software (https://maxquant.net/perseus/) (but for *Windows PC only*…!). Tutorials on MaxQuant and Perseus can be found on a YouTube channel (https://www.youtube.com/c/MaxQuantChannel/videos)
- Annotation exists for many types of biological knowledge. It can be imported and used as fit : GO, KEGG, REACTOME, CORUM, Flybase, Interpro, Pfam, Prosite, SGD, WormBase, UniProt,…
- Many **R** packages used for genomics data analysis can be applied to proteomics data after raw MS data processing (e.g. normalization, annotation enrichment, ..). Dedicated packages also exist.

Appendix

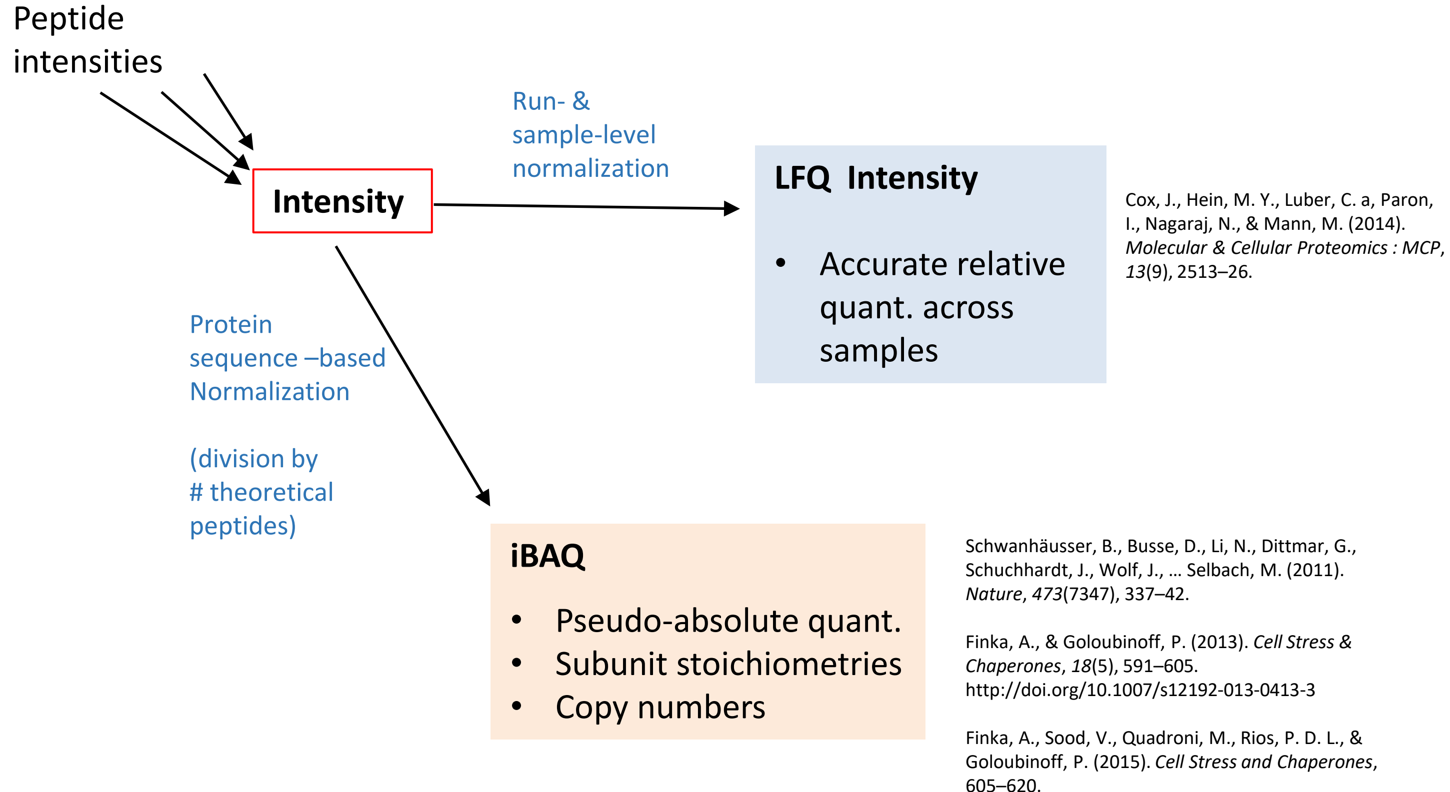
1. Additional information on the quantitation variables produced from MaxQuant for different pipelines

MaxQuant output variables (protein level)



*** : main output**

MaxQuant label-free output variables (protein level)



Thank you for your attention

We hope it was useful…..