UNIL | Université de Lausanne

Protein Analysis Facility

Proteome quantitation by Mass Spectrometry (MS)

A short guide to data processing and interpretation

Version 5 – October 2024

GOALS

EXPLAIN

The final part of the data analysis pipeline

DESCRIBE

The format of data

PROVIDE

Options for downstream data processing and interpretation

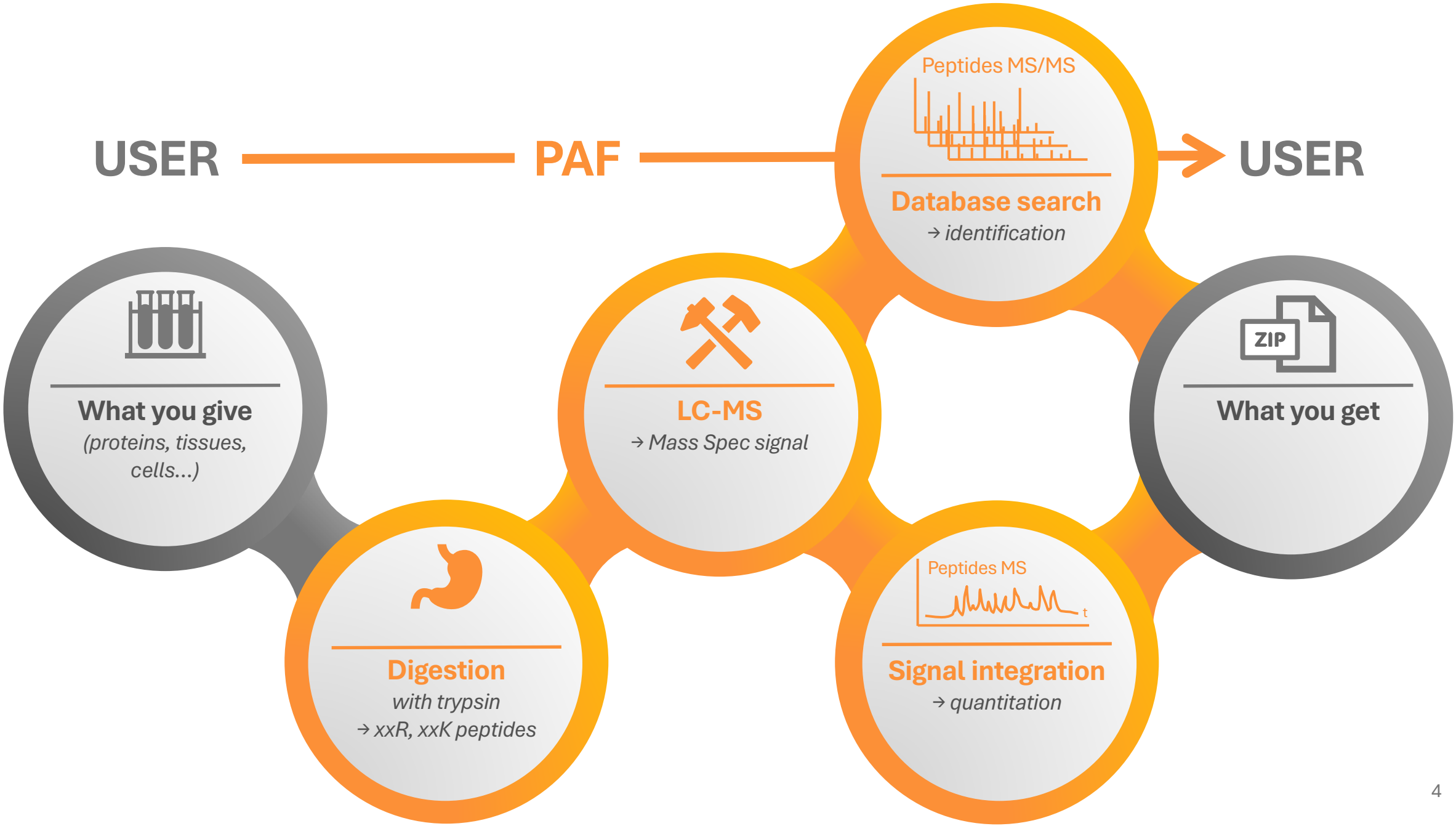
ANSWER

Frequently asked questions

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

INDEX

PROTEOMICS WORKFLOW



THINGS TO REMEMBER Before going further

Proteins are identified, **NOT** *de novo* sequenced

A protein present **in the sample** but **not in the database** will not be identified

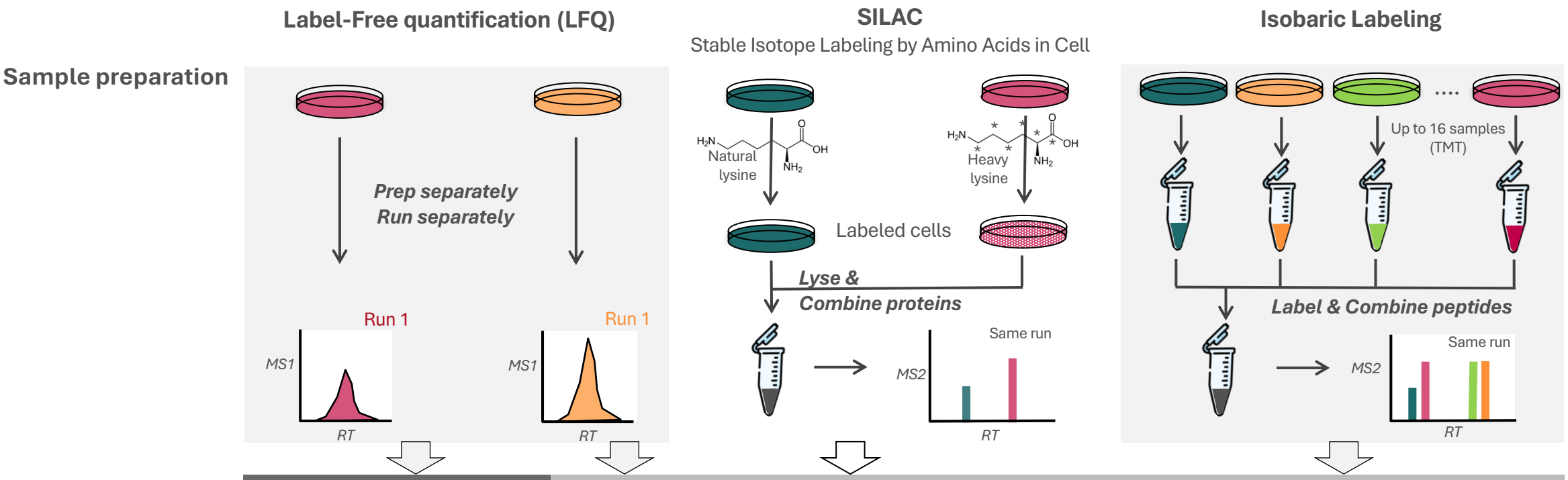
The choice of **database** is critically important

***Not** using a **model organism** ( , ) or looking for **modified sequences** such as tagged proteins?*

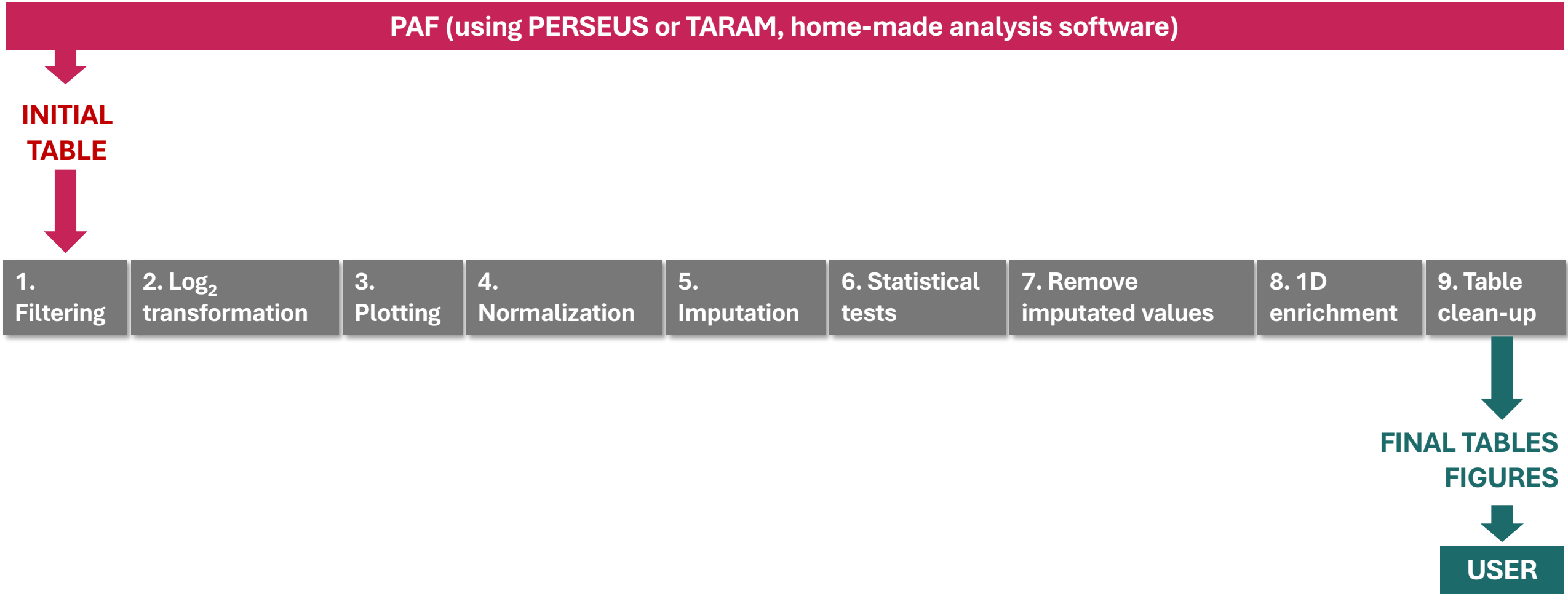
Please provide the relevant database or the related sequences

DIFFERENT MS PIPELINES, SAME TYPE OF OUTPUT

Shuken (2023) Journal of Proteome Research 22, 2151-2171



DATA ANALYSIS Overview



To remove irrelevant data such as

Reverse hits

In large datasets, an additional search is performed against a «decoy» database composed of reverse sequences to determine the frequency of false positives. Scoring parameters are then adjusted to have 1% FDR (a.k.a 1% decoy hits) in the final results.

Contaminants

Keratins, trypsin, bovine serum albumin …
They are present in samples because of the experiment or as common environmental contaminations

Only identified by site

Protein groups identified only by a modified peptide

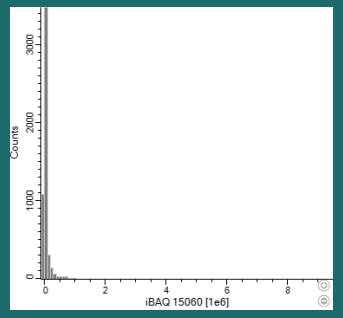
Filter on valid values

Protein groups that have many missing values cannot be quantified with confidence

To normalize the distribution and facilitate plotting

LINEAR (raw) data

Vast dynamic range
Asymmetric distributions



Multiplication
Division
0

FOLD change
A/B = X

Examples:
A/B = 2
A/B = 0.5

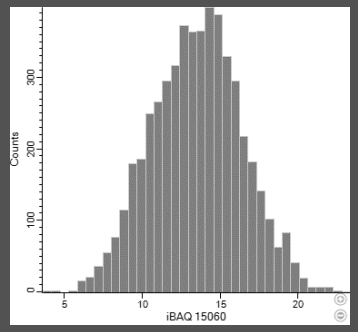
Sum
Subtraction
NaN (Not A Number)

log_2 difference
$log_2(A) - log_2(B) = log_2(X)$

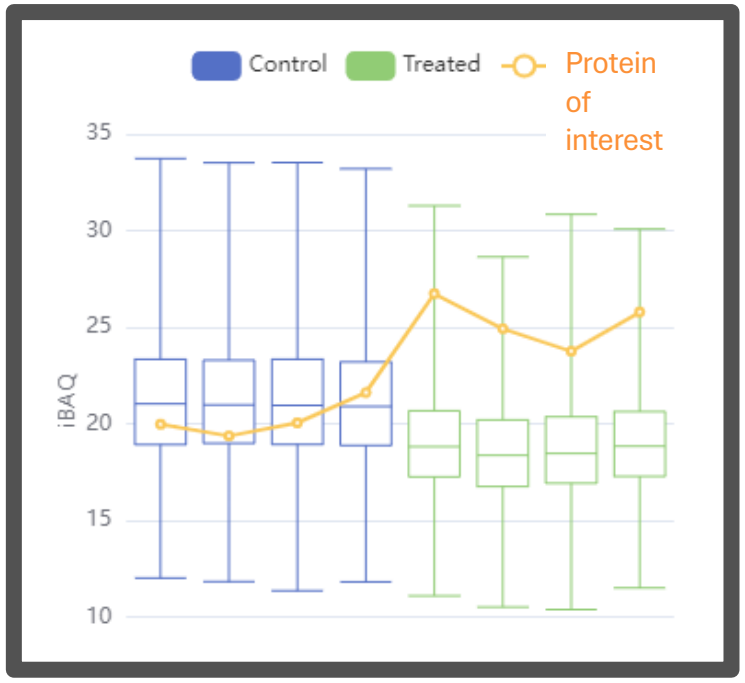
Examples:
$log_2(A) - log_2(B) = 1$
$log_2(A) - log_2(B) = -1$

LOG (transformed) data

Manageable dynamic range (plots)
More normal distributions



To look at data globally



Medians Distributions

Detect loading differences

Detect measurement biases

Negative & positive controls

Assess sensitivity

Biological Quality Control

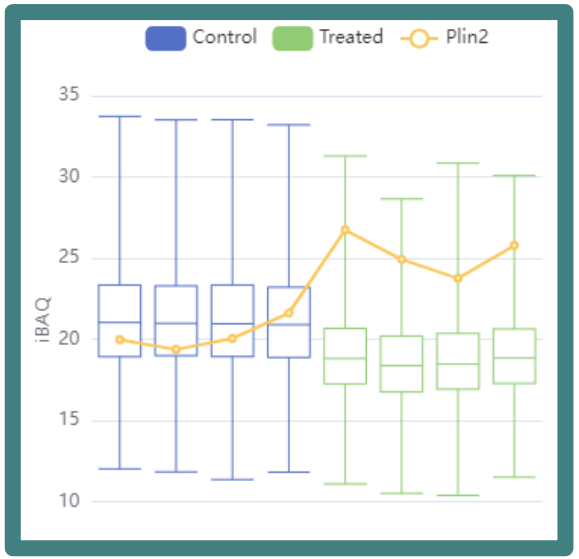
Correlation between replicates / Clustering

Quality Control of conditions

Look for outliers, emerging groups

To compensate for global differences in sample amount

RAW DATA



ASSUMPTION

Most of the proteome does not change
(stable background)

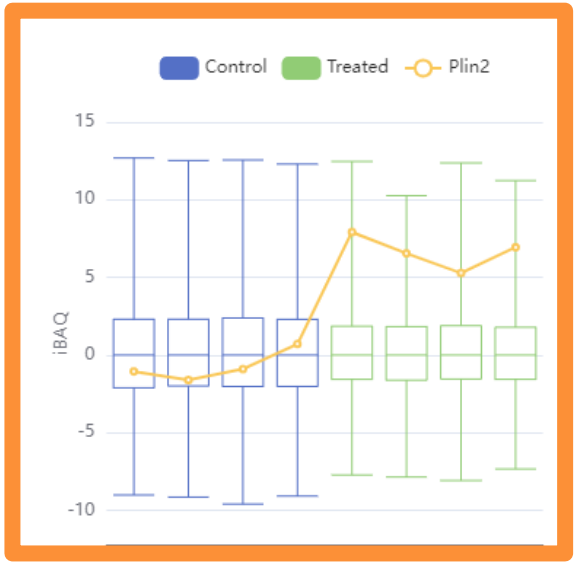
METHODS

Median subtraction is the most conservative normalization.

For IPs, data could be normalized based on intensity of a protein of interest BUT this may skew some datasets and result in false positives.

Some data such as MaxQuant LFQ or Spectronaut Quantity are already normalized.

NORMALIZED DATA



To deal with missing values

1 Quantitative proteomics data always contain **missing values**

2 Missing values could come from :

- stochasticity of data acquisition
- peptides/proteins near/below the limit of detection
- truly absent proteins

3 It is impossible to discriminate between the events in point 2.

4 **MISSING VALUES**

"0" in linear data
"NaN" in log data (Not A Number)

5 Some types of data contain more missing values (LFQ, iBAQ) than others (TMT, DIA)

6 **Missing values** are a data analysis problem.

Fold changes cannot be calculated.
Proteins cannot be plotted.

7 **Imputation**

= substitution of missing values with values used for computation

8 Imputed values are often chosen at the low end of the distribution of the values, assuming that the reason for the missing values is very low signal intensity.

A (very) low fixed value could also be used in some cases.

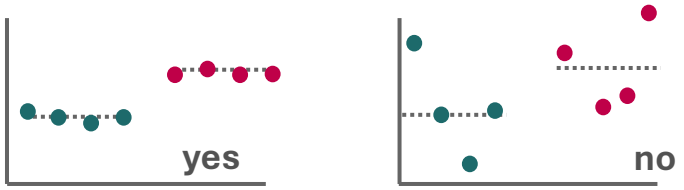
9 **Imputation has to be applied with caution and on a limited scale**

Imputed data should be flagged to be easily identified (and removed) later.

To determine whether the observed differences are statistically significant

To pass the test, difference between groups must be larger than the variability within groups

Example



Statistical test

Most frequent tests :
Student's T-test or Welch Test

Statistical power

= Number of replicates (crucial !)
N = 3 is the minimum

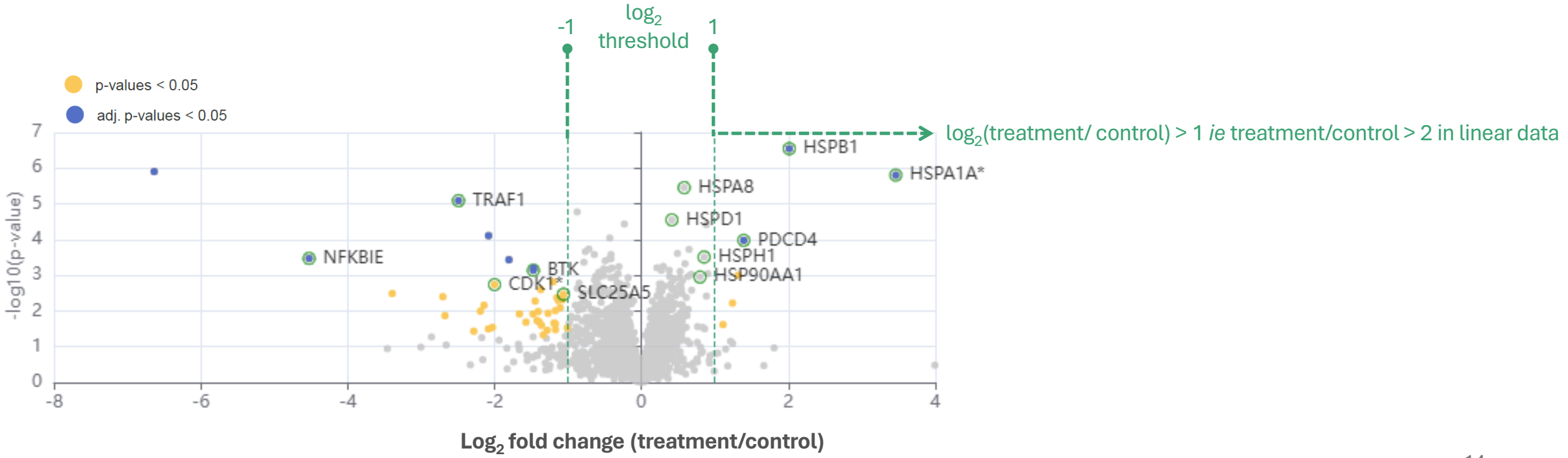
Multiple testing correction

Necessary when working with large datasets
Exemple of method : Benjamini-Hochberg

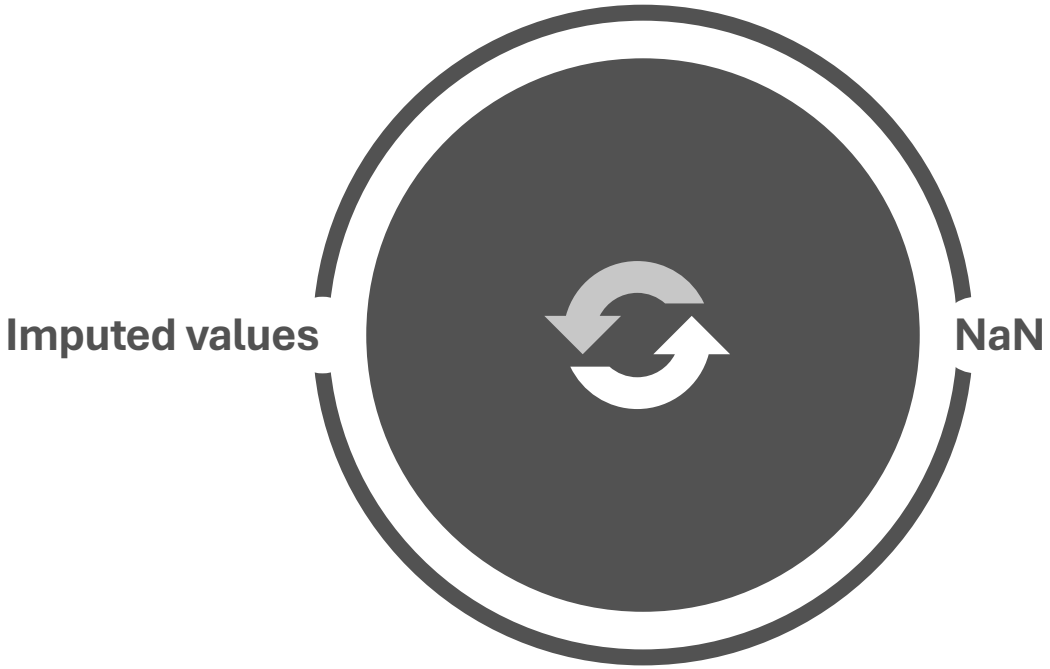
Correction of p-value yields adjusted p-value

Volcano plot is a common way to represent statistical significance (*y-axis*) and fold change (*x-axis*) in a single plot.
By using –log10 transformation, low p-values (highly significant) appear toward the top of the plot.

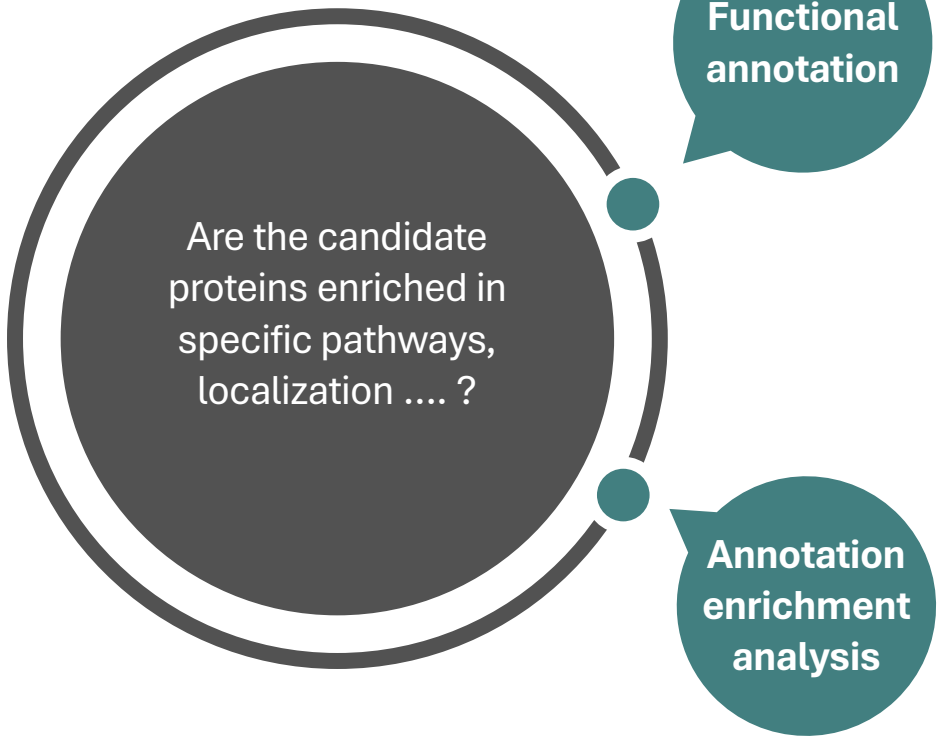
Each dot is a protein group.
The gene name of particulary interesting ones is provided.
Significant protein groups are highlighted using a color-coded system.



The final dataset should contain only real measured values



For functional interpretation of the data



Terms describing functional or structural characteristics, interactions, pathways or other features are added to the protein groups.

Sources (not exhaustive):

Gene Ontology (GO)	http://geneontology.org/
KEGG	https://www.genome.jp/kegg/
Corum	https://mips.helmholtz-muenchen.de/corum/
Reactome	https://reactome.org/

A statistical approach is used to identify annotations that are significantly over-represented in association with a quantitative trait, typically the fold-change between 2 conditions.

Limitations :

Statistical bias since the different areas of biology are more or less well annotated and/or curated

Redundant and complex annotations

To get a final dataset

Columns are removed and reordered

A table filtered for proteins identified by 3 or more peptides can be provided (more reliable quantitative values)

ZIP file with report, tables and plots.

Visit our website for an exemple :
https://wp.unil.ch/paf/useful-infos/

FINAL TABLE FORMAT The exact names of the columns can vary

Difference of means of conditions.
Equivalent to fold change, but in log_2 scale

Includes shared peptides but assigned only to highest scoring protein group (= razor peptides).
Peptides used for quantitation.

Adj. p-val = p-value after multiple testing correction

Is protein group passing statistical test ?
Usually with adj p-val<0.05

Total number of peptides used for identification/quantitation.
Includes shared peptides

Unique peptides for this protein group

t-test p-value (uncorrected)

Quantitative values
Could be iBAQ, LFQ, intensity, quantity

Ctrl R1 iBAQ.log2	Ctrl R2 iBAQ.log2	Ctrl R3 iBAQ.log2	Treated R1 iBAQ.log2	Treated R2 iBAQ.log2	Treated R2 iBAQ.log2	iBAQ	Majority.protein.IDs	Protein.names	Gene.names	Fasta.headers	Peptides	Razor.unique.peptides	Unique.peptides	Ctrl R1.Razor.unique.peptides	Ctrl R2.Razor.unique.peptides	Ctrl R3.Razor.unique.peptides	Treated R1.Razor.unique.peptides	Treated R2.Razor.unique.peptides	Treated R3.Razor.unique.peptides	Mol.weight.kDa	Sequence.coverage	Score	Intensity	MS.MS.count	p.value.Treated-Control	adj. p.value.Treated-Control	log2.fold.change.Treated-Control	is.significant.Treated-Control	GOBP name	GOMF name	GOCC name	KEGG name
3.52	3.44	3.29	NaN	NaN	NaN	2.37E+08	P63328	Serine/threonine-protein p	Ppp3ca	sp|P63328|PP2	22	22	14	15	14	14	0	0	0	58.6	36.3	159.18	1.34E+10	269	NaN	NaN	NaN	false	G1/S	cataly	nucle	MAP
5.56	5.71	5.61	-0.30	-1.12	1.66	1.66E+09	P63330	Serine/threonine-protein p	Ppp2ca	sp|P63330|PP2	21	21	4	15	15	14	1	1	1	35.6	70.6	323.31	5.24E+10	580	8.3E-05	2.0E-04	-5.7	true	mitc	G-pro	protei	mRN
6.22	6.54	6.03	3.39	2.17	1.50	3.22E+09	P67778;Q5SQG5	Prohibitin	Phb	sp|P67778|PHI	27	27	27	14	16	16	12	7	7	29.8	73.9	323.31	1.45E+11	637	2.3E-04	4.5E-04	-3.7	true	neg	protei	nucleus;n	
3.52	3.93	3.74	NaN	-1.2	NaN	3.78E+08	P67871;G3UZJ5;G3UZA	Casein kinase II subunit be	Csnk2b	sp|P67871|CSK	9	9	9	8	8	8	0	1	0	24.9	59.5	153.91	1.06E+10	228	1.4E-04	3.0E-04	-4.6	true	mor	chrom	chron	Ribo
7.53	7.80	7.75	4.5	5.41	5.23	5.53E+09	P67984	60S ribosomal protein L22	Rpl22	sp|P67984|RL2	9	9	8	7	5	5	3	3	4	14.8	39.1	153.9	6.67E+10	222	2.0E-05	7.1E-05	-2.6	true	cell	patter	nucle	Ribo
7.46	7.62	7.49	1.96	1.35	2.18	4.53E+09	P68040	Guanine nucleotide-bindir	Gnb2l1	sp|P68040|RAC	26	26	26	20	18	18	6	6	7	35.1	78.5	323.31	1.50E+11	794	7.6E-08	2.8E-06	-5.8	true	regu	enzym	phag	Mea
-0.28	-0.60	-1.10	NaN	NaN	NaN	2.85E+07	P68181	cAMP-dependent protein k	Prkacb	sp|P68181|KAI	17	4	4	2	1	1	0	0	0	40.7	31.6	12.47	1.02E+09	44	NaN	NaN	NaN	false	neu	nucle	nucle	MAP
7.25	7.31	7.39	2.35	2.00	2.91	3.99E+09	P68254	14-3-3 protein theta	Ywhaq	sp|P68254|143	26	22	22	14	13	14	7	6	9	27.8	64.1	323.31	2.14E+11	735	3.2E-07	5.7E-06	-4.8	true	regu	bindi	cytop	Cell

LOG2 scale !
Normalized values can be negative ie lower than median

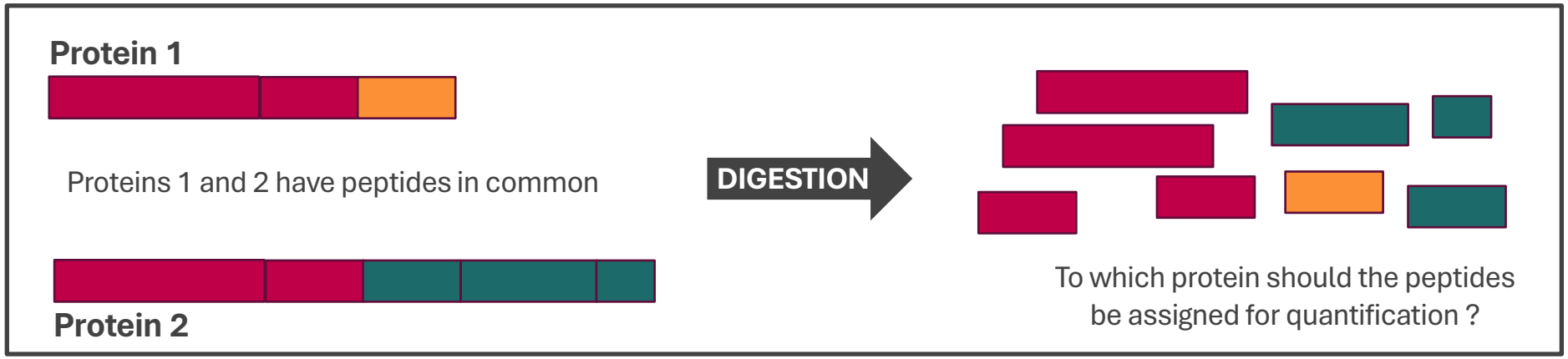
Database identifier.
First one is the best

Use this for filtering

Total number of matched spectra.
Can also be used for filtering/validation if column is present (SILAC or TMT)

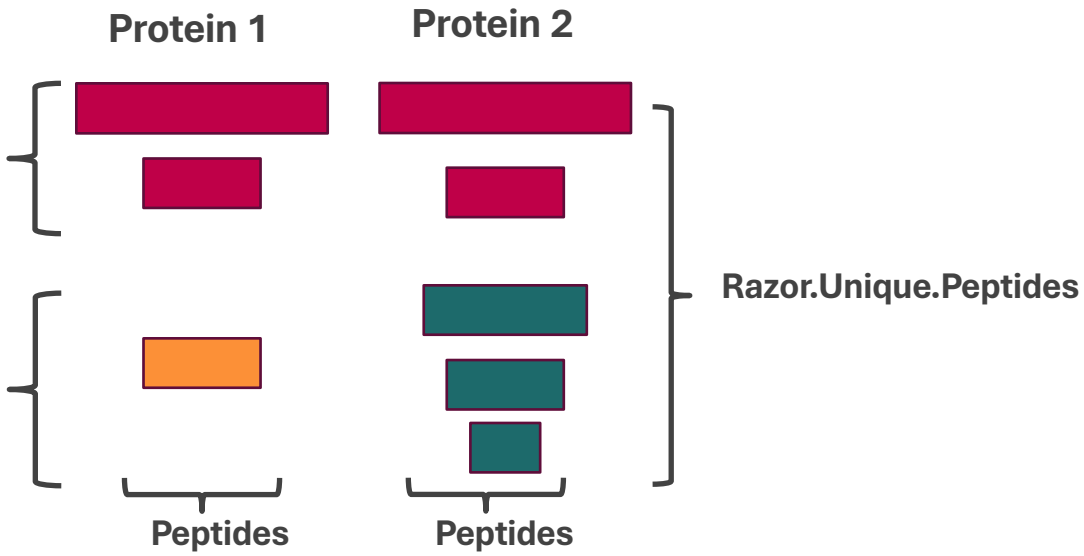
Functional annotations
(many terms per line)

SOME DEFINITIONS Unique and razor peptides

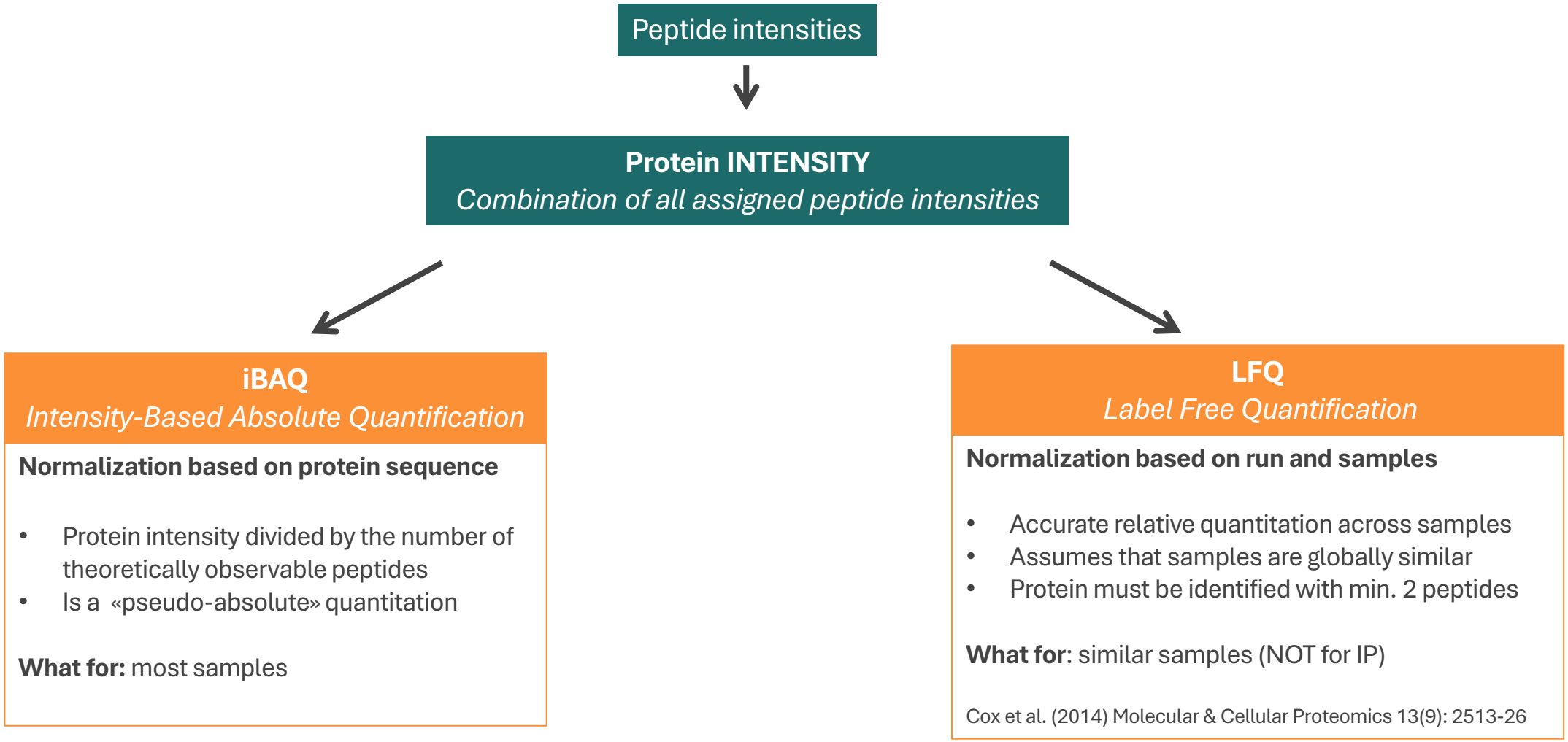


Razor peptides
= non-unique peptides usually assigned to the protein group with the highest number of peptides
Protein 2 in this case

Unique peptides



MORE DEFINITIONS MaxQuant Output variables



Note : Spectronaut calculates a «quantities», which are mostly similar to MaxQuant LFQ values

GENERAL FACTS AND THINGS TO REMEMBER

Proteomes

tend to change less than transcriptomes

The amplitudes of changes are often smaller and the changes slower.

Statistical power

is crucial to find significantly different proteins

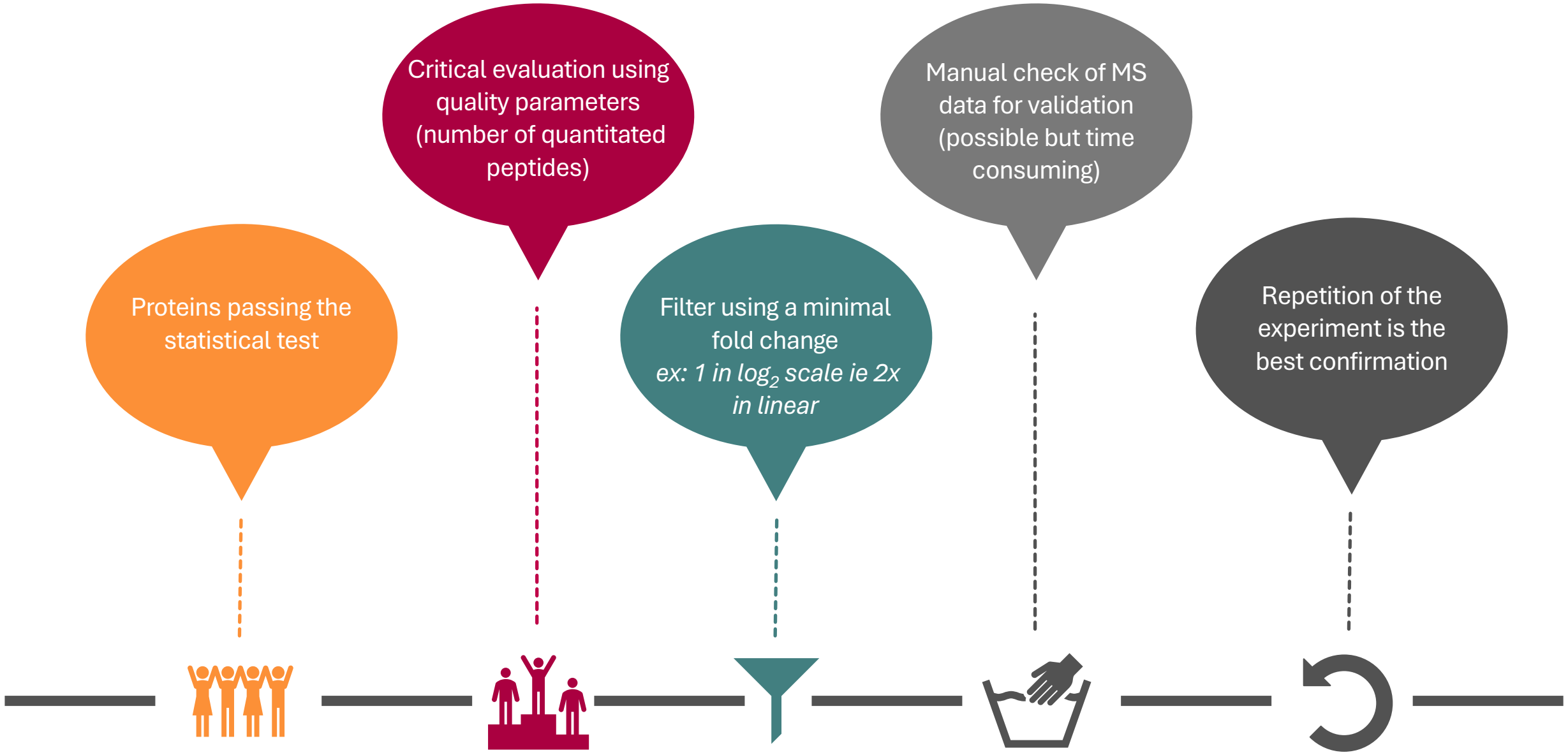
Power is mostly determined by the number N of biological replicates.
N=3 is the minimum for reliable (and publishable !) data.
N >3 is recommended when quantifying mild changes, in case of high variability or for studies on individual animals.

Quantitation reliability

is lower for weak signals and proteins with a low peptide count

These are mostly low abundance proteins.

EVALUATION OF PROCESSED OUTPUT How to define a list of candidates



...OUPS, IT DID NOT WORK OUT

Sometimes, no statistically significant proteins emerge

It does not necessarily mean there is no difference

How to move on ?

Most frequent reasons:

Small differences between conditions
Insufficient statistical power (**N** too low)
High variability within groups

Often more than one of these factors combined !

It may simply indicate that the experiment was not perfect

- Select the best candidates (using the best fold-change & p-value, proteins identified with minimum 3 peptides)
- Try to quantify in a targeted way (WB, FACS, ELISA)
- Look at annotation enrichment for global trends
- Repeat with more replicates

PUBLICATION AND DATA SUBMISSION TO REPOSITORIES

Please read our guidelines on **acknowledging core facility work and authorship**

https://wp.unil.ch/paf/policy/

Remember to ask for **Material & Methods** descriptions sooner rather than later.

Retrieving all relevant information about your experiment may become challenging after a few years.

Upload of raw and processed data **to repositories** is highly recommended or mandatory for most journals.
We recommend **proteomexchange.org** (part of PRIDE https://www.ebi.ac.uk/pride/).

We handle data preparation and upload for our users. However, we require **at least 2 weeks** prior to manuscript submission.
For really complex datasets, additional time may be required. We may need to charge you for the work involved.

Submission to the data repository requires

- Raw MS data
- Project title, description
- Materials and Methods
- Protein ID and quant data

A template for submission to PRIDE/Proteomexchange is available on :
https://wp.unil.ch/paf/useful-infos/

INFORMATION RESOURCES

MaxQuant

https://maxquant.net/maxquant/

MaxQuant and Perseus are free

but for *Windows or Linux only* !

Tutorials can be found on YouTube

(https://www.youtube.com/c/MaxQuantChannel/videos)

Perseus

https://maxquant.net/perseus/

Spectronaut

https://biognosys.com/software/spectronaut/

Many **R** packages used for genomics data analysis can be applied to proteomics data after raw MS data processing (e.g. for normalization, annotation enrichment, ..).

Dedicated packages also exist.

FREQUENTLY ASKED QUESTIONS general considerations

Why are you talking about protein groups and not proteins ?

"Some proteins cannot be unambiguously identified by unique peptides (but have only shared peptides).
They are grouped in one protein group and quantified together, e.g. if all detected peptides of protein A also belong to protein B, A and B form one protein group." (Maxquant)

Why should I do replicates ?

The question should rather be "why did you not do replicates?"

Good science must be reproducible. Replicates provide « an estimate of between samples error». They also enhance reliability for low-abundance protein groups or subtle differences between samples.

What is the point of doing a pilot experiment ?

To test and optimize the different steps of both your and our experimental designs (such as proteins extraction and digestion from our side). A pilot gives an idea about feasibility and can save everyone time and money.

Can I analyse together data from replicates that were generated and analysed at different moment ?

Technically yes but there will be batch effects. By far the best consistency is obtained when all samples to be compared are prepared and analysed in the same session.

FREQUENTLY ASKED QUESTIONS about imputation

How does imputation work ?

Please read page 12 of this manual

Is imputation random?

Yes, imputation is usually done with random values taken from a low-shifted distribution, to simulate “real” data for low abundance proteins.
A fixed low value can be used but this can skew the results of statistical tests, as they assume “normal” data distributions. Fixed values are used only when N=1.

Is imputation reproducible?

Not 100%, since the values used are random-generated. If the imputation is repeated, the imputed values will be slightly different.

Imputation is random and not reproducible : how is it reliable ?

Since the values are all low and in a small range, the impact of the randomness of the imputed data on the final results should be limited.
We also do not perform calculations on totally imputed data, i.e. when comparing two groups, at least in one group there should be real values.

FREQUENTLY ASKED QUESTIONS about data filtering

How do I identify outlier samples ?

The PCA (principal component analysis) plot is the best tool to spot outliers. It also enables us to evaluate sample homogeneity within groups and differences between conditions.

Why are data filtered using peptides number ?

We assume that protein groups identified with few peptides may represent low-abundance proteins / may be less reliable.

We typically prioritize protein groups identified with minimum 2 Razor.unique peptides (i.e. >1 razor.unique peptides).

Why data are filtered using a certain number of valid values into a group (when replicates are present)?

Assumption: protein groups identified in some, but not all, replicates of a group may represent low-abundance proteins / may be less reliable.
We typically prioritize protein groups that are consistently identified across all replicates within at least one sample group.
Ex: protein groups with at least 3 intensity values in one group when working with triplicates

What is the minimum (fold) change that I can consider as "real" ?

As a rule of thumb, fold changes smaller than +/- 0.5 in $\log_2$ should not be considered.
A safe minimal threshold is +/- 1 $\log_2$, a two-fold change in linear data.
Annotation enrichment analysis is a way to highlight global changes that are of low magnitude.

APPENDIX

Additional information about annotation enrichment

ANNOTATION ENRICHMENT Basic concepts

Annotation

Terms are added to each protein in the list of interest.

ex: GO terms

Sorting

The list is sorted based on a relevant experimental parameter.

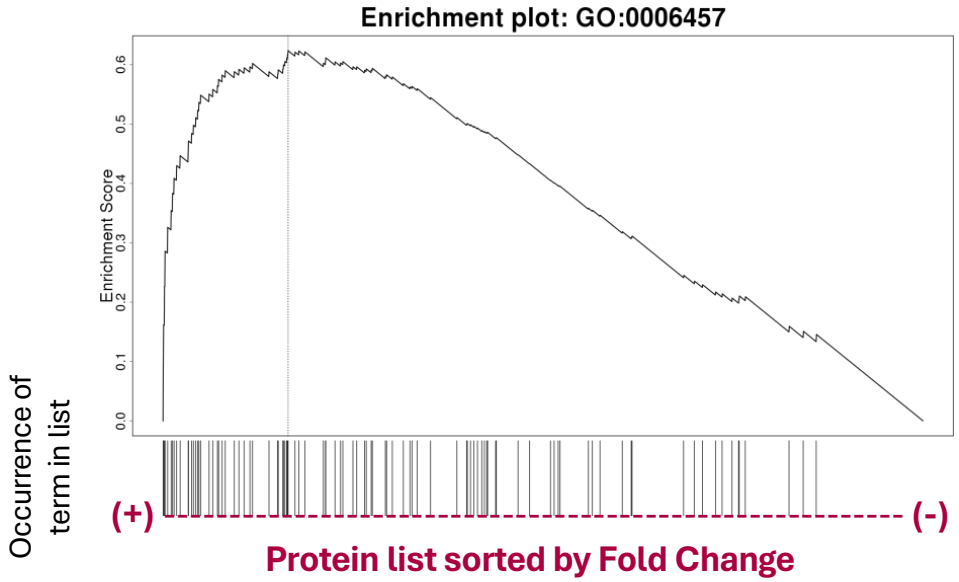
ex: fold change between the conditions

Statistics

An algorithm is applied to assess whether each annotation term is significantly enriched at either the top or bottom of the list, indicating a tendency to be associated with proteins that either increase or decrease between conditions.

Correction

A p-value is calculated and should be corrected for multiple testing in case of large datasets with many annotation terms.



Note: the algorithm used in our analyses is described in Cox & Mann (2012) *BMC Bioinformatics*,*13* (*Suppl 16), S12*. https://doi.org/10.1186/1471-2105-13-S16-S12.
It is conceptually similar to the widely used GSEA algorithm.

ANNOTATION ENRICHMENT Result table

Parameter used for sorting the table. Usually $\log_2$ (fold change)

P-value for enrichment of this term

Number of protein groups in the list with this annotation type

Corrected p-value (Benjamini-Hochberg method). A cut-off of 0.02 is usually applied to generate this list.

Column	Type	Name	Size	Score	P value	Benj. Hoch. FDR	Mean	Median
log2.fold.change.Treated_Control	GOBP name	drug transmembrane tr	6	0.92	1.01E-04	0.005	3.315	3.83
log2.fold.change.Treated_Control	GOMF name	polyamine transmemb	8	0.74	2.71E-04	0.012	3.284	3.66
log2.fold.change.Treated_Control	GOBP name	drug transport	7	0.89	4.68E-05	0.003	2.997	3.60
log2.fold.change.Treated_Control	GOMF name	drug transmembrane tr	7	0.89	4.68E-05	0.002	2.997	3.60
log2.fold.change.Treated_Control	GOBP name	mannose transport	7	0.96	1.02E-05	0.001	3.781	3.50
log2.fold.change.Treated_Control	GOMF name	fructose transmembrar	7	0.96	1.02E-05	0.001	3.781	3.50
log2.fold.change.Treated_Control	GOMF name	iron ion transmembrar	6	0.84	3.51E-04	0.015	2.313	2.07
log2.fold.change.Treated_Control	GOCC name	intrinsic to plasma me	24	0.56	2.68E-06	0.000	1.971	1.97
log2.fold.change.Treated_Control	GOCC name	anchored to membrane	10	0.77	2.86E-05	0.001	1.824	1.97
log2.fold.change.Treated_Control	GOBP name	copper ion transport	10	0.74	5.34E-05	0.003	2.085	1.95
log2.fold.change.Treated_Control	GOCC name	integral to plasma mer	21	0.58	4.43E-06	0.000	2.071	1.91
log2.fold.change.Treated_Control	GOCC name	vacuolar lumen	6	0.84	3.97E-04	0.007	2.646	1.70
log2.fold.change.Treated_Control	GOMF name	cofactor transporter ac	8	0.77	1.64E-04	0.008	1.926	1.65
log2.fold.change.Treated_Control	GOMF name	amine transmembrane	32	0.49	2.23E-06	0.000	1.522	1.56
log2.fold.change.Treated_Control	GOBP name	cofactor transport	10	0.67	2.61E-04	0.010	1.542	1.52

GOBP : Biological process
GOCC: Cellular Compartment
GOMF: Molecular fonction

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

Can be seen as normalized version of mean/medians. Values are always between -1 and +1, centered at 0.

Mean resp. Median of «T-test differences» for each GO term, calculated on the protein groups considered (Size). In $\log_2$, >0 means increasing in Treated and *vice versa*.

ANNOTATION ENRICHMENT

How to retrieve proteins associated with a specific GO annotation ?

Column	Type	Name	Size	Score	P value	Benj. Hoch. FDR	Mean	Median
log2.fold.change.Treated_Control	GOBP name	drug transmembrane tr	6	0.92	1.01E-04	0.005	3.315	3.83
log2.fold.change.Treated_Control	GOMF name	polyamine transmemb	8	0.74	2.71E-04	0.012	3.284	3.66
log2.fold.change.Treated_Control	GOBP name	drug transport	7	0.89	4.68E-05	0.003	2.997	3.60
log2.fold.change.Treated_Control	GOMF name	drug transmembrane tr	7	0.89	4.68E-05	0.002	2.997	3.60
log2.fold.change.Treated_Control	GOBP name	mannose transport	7	0.96	1.02E-05	0.001	3.781	3.50
log2.fold.change.Treated_Control	GOMF name	fructose transmembrar	7	0.96	1.02E-05	0.001	3.781	3.50
log2.fold.change.Treated_Control	GOMF name	iron ion transmembrar	6	0.84	3.51E-04	0.015	2.313	2.07
log2.fold.change.Treated_Control	GOCC name	intrinsic to plasma mer	24	0.56	2.68E-06	0.000	1.971	1.97
log2.fold.change.Treated_Control	GOCC name	anchored to membrane	10	0.77	2.86E-05	0.001	1.824	1.97
log2.fold.change.Treated_Control	GOBP name	copper ion transport	10	0.74	5.34E-05	0.003	2.085	1.95
log2.fold.change.Treated_Control	GOCC name	integral to plasma mer	21	0.58	4.43E-06	0.000	2.071	1.91
log2.fold.change.Treated_Control	GOCC name	vacuolar lumen	8	0.84	3.97E-04	0.007	2.646	1.70
log2.fold.change.Treated_Control	GOMF name	cofactor transporter ac	8	0.77	1.64E-04	0.008	1.926	1.65
log2.fold.change.Treated_Control	GOMF name	amine transmembrane	32	0.49	2.23E-06	0.000	1.522	1.56
log2.fold.change.Treated_Control	GOBP name	cofactor transport	10	0.67	2.61E-04	0.010	1.542	1.52
log2.fold.change.Treated_Control	GOCC name	fungal-type vacuole lur	5	0.83	0.00133107	0.020	2.784	1.45
log2.fold.change.Treated_Control	GOBP name	transition metal ion tra	29	0.51	1.85E-06	0.000	1.359	1.40
log2.fold.change.Treated_Control	GOBP name	iron ion transport	11	0.62	4.15E-04	0.014	1.577	1.40
log2.fold.change.Treated_Control	GOCC name	organellar large riboso	41	0.67	1.16E-13	0.000	1.168	1.33
log2.fold.change.Treated_Control	GOCC name	mitochondrial large rib	41	0.67	1.16E-13	0.000	1.168	1.33
log2.fold.change.Treated_Control	GOMF name	amino acid transmemb	27	0.40	3.62E-04	0.015	1.107	1.29

ANNOTATION ENRICHMENT

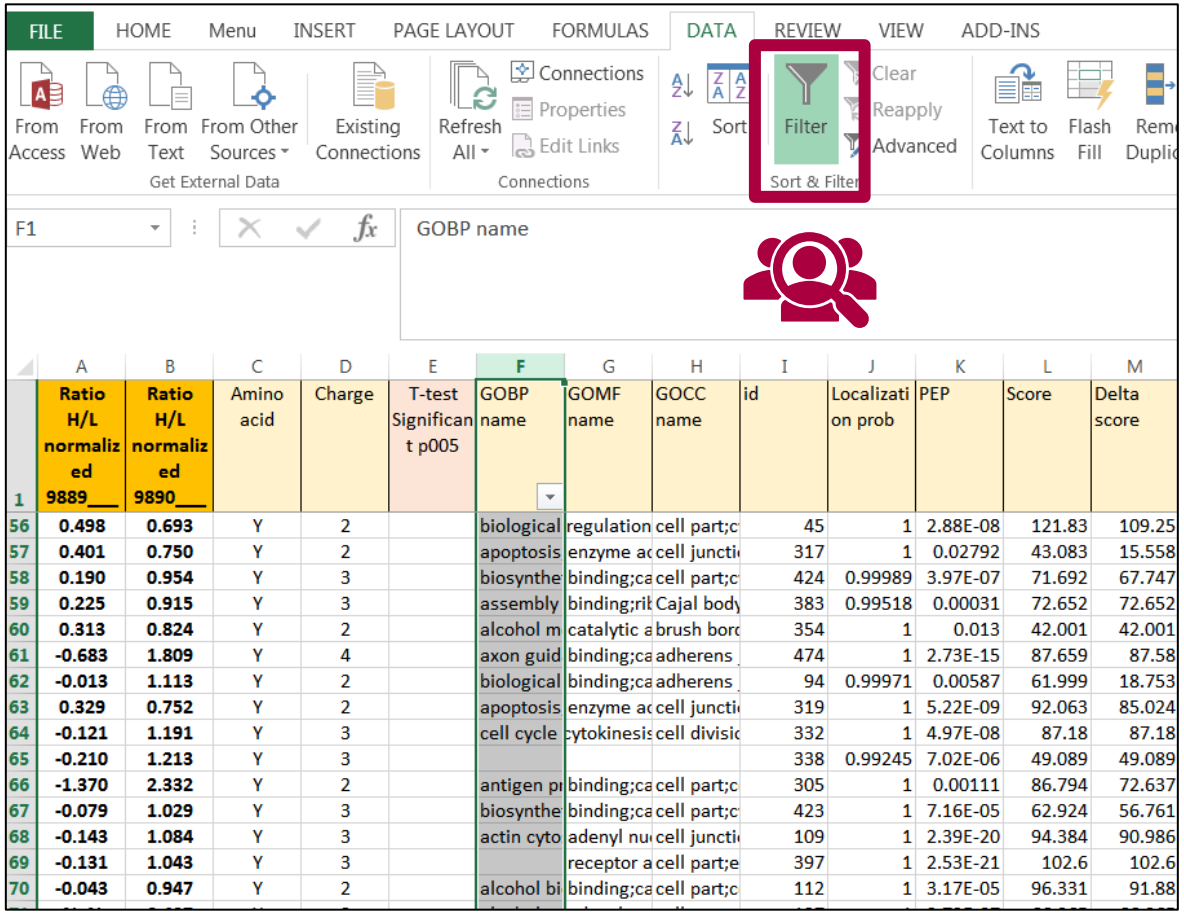
How to retrieve proteins associated with a specific GO annotation ?

Open the table in Excel

Select the column that contains the annotation of interest

Ex: GOBP

Go to «data» and click «Filter»



1	A Ratio H/L normalized 9889___	B Ratio H/L normalized 9890___	C Amino acid	D Charge	E T-test Significant p005	F GOBP name	G GOMF name	H GOCC name	I id	J Localization prob	K PEP	L Score	M Delta score
56	0.498	0.693	Y	2		biological	regulation	cell part;c	45	1	2.88E-08	121.83	109.25
57	0.401	0.750	Y	2		apoptosis	enzyme ac	cell junctio	317	1	0.02792	43.083	15.558
58	0.190	0.954	Y	3		biosynthe	binding;ca	cell part;c	424	0.99989	3.97E-07	71.692	67.747
59	0.225	0.915	Y	3		assembly	binding;rib	Cajal body	383	0.99518	0.00031	72.652	72.652
60	0.313	0.824	Y	2		alcohol m	catalytic a	brush bord	354	1	0.013	42.001	42.001
61	-0.683	1.809	Y	4		axon guid	binding;ca	adherens	474	1	2.73E-15	87.659	87.58
62	-0.013	1.113	Y	2		biological	binding;ca	adherens	94	0.99971	0.00587	61.999	18.753
63	0.329	0.752	Y	2		apoptosis	enzyme ac	cell junctio	319	1	5.22E-09	92.063	85.024
64	-0.121	1.191	Y	3		cell cycle	cytokinesis	cell divisio	332	1	4.97E-08	87.18	87.18
65	-0.210	1.213	Y	3					338	0.99245	7.02E-06	49.089	49.089
66	-1.370	2.332	Y	2		antigen pr	binding;ca	cell part;c	305	1	0.00111	86.794	72.637
67	-0.079	1.029	Y	3		biosynthe	binding;ca	cell part;c	423	1	7.16E-05	62.924	56.761
68	-0.143	1.084	Y	3		actin cyto	adenyl nu	cell junctio	109	1	2.39E-20	94.384	90.986
69	-0.131	1.043	Y	3			receptor a	cell part;e	397	1	2.53E-21	102.6	102.6
70	-0.043	0.947	Y	2		alcohol bi	binding;ca	cell part;c	112	1	3.17E-05	96.331	91.88

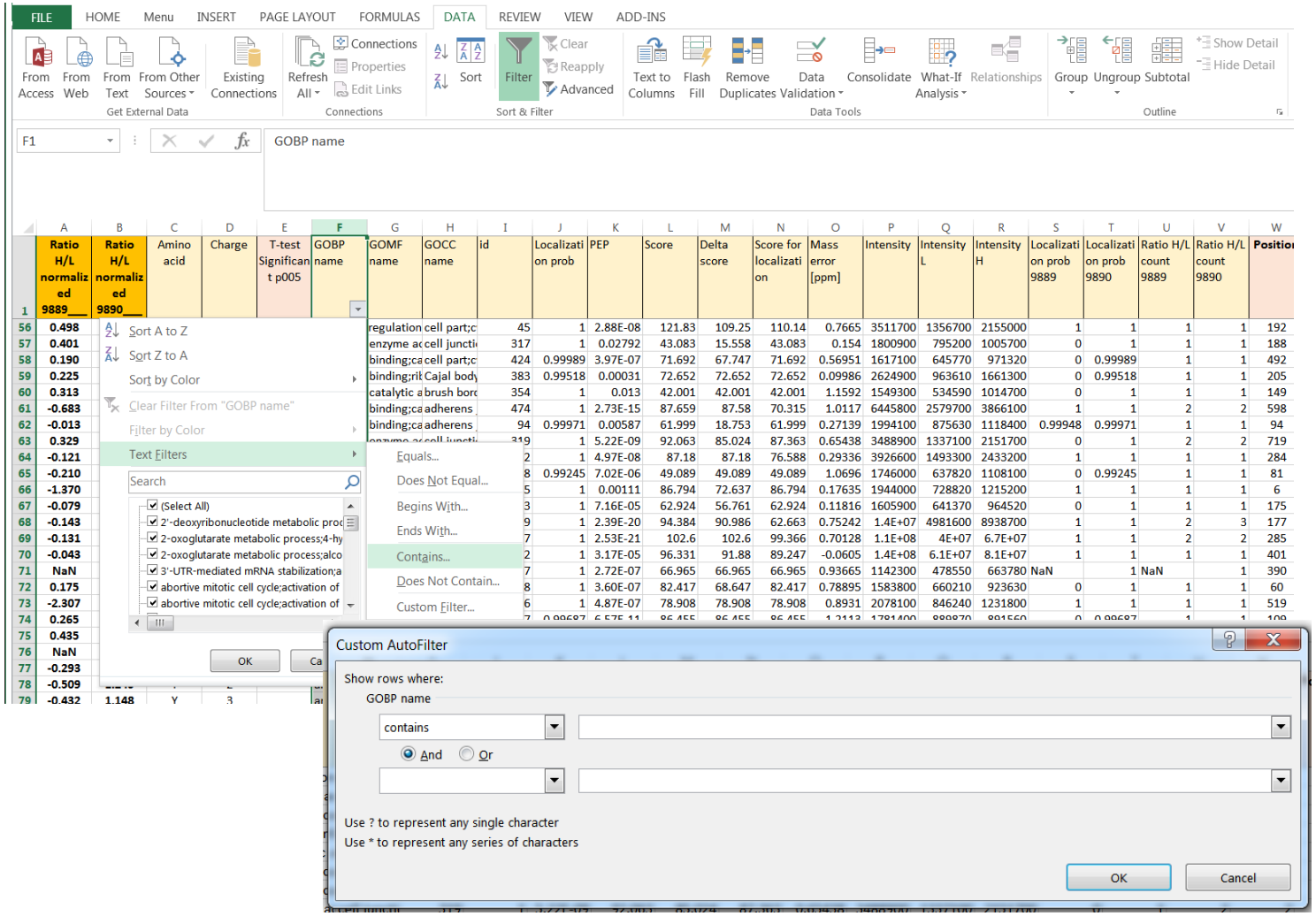
ANNOTATION ENRICHMENT

How to retrieve proteins associated with a specific GO annotation ?

Go to the pull-down menu

Select «text filters» and «contains»

Enter the name of the annotation you are searching for in the «contains» field




Thank you for your attention
We hope it was useful.....