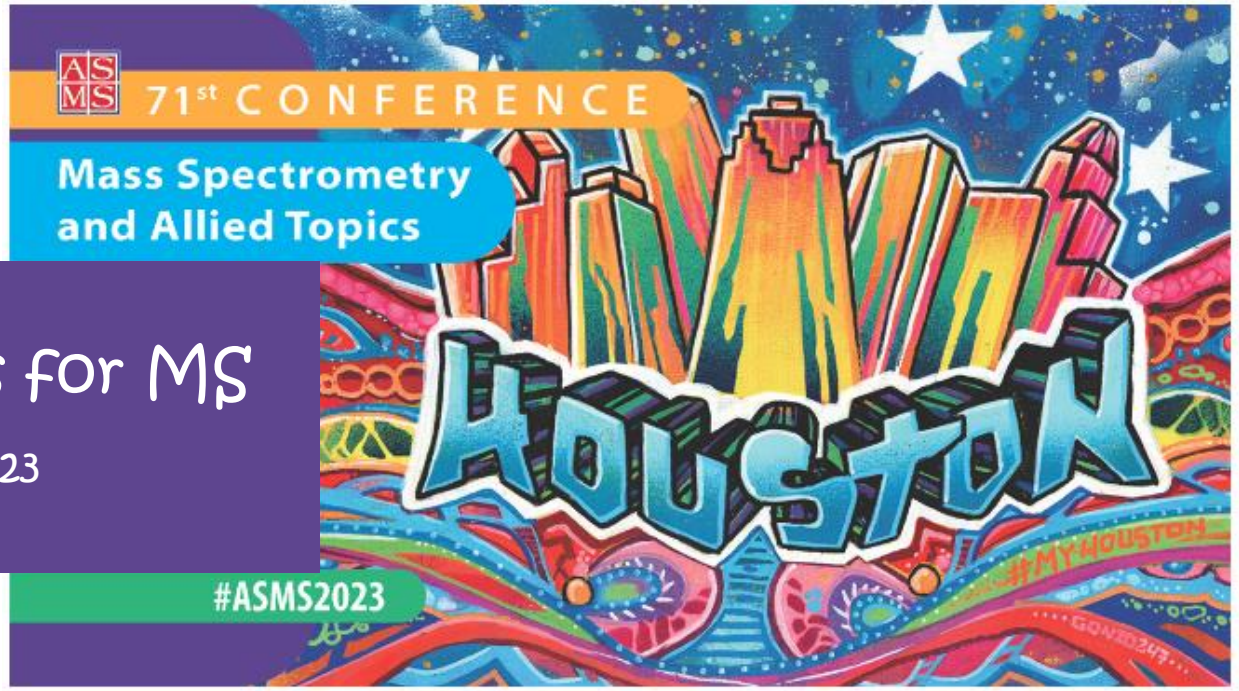
Bioinformatics for MS

June 6th, 2023



From data to biology: using -omics datasets to generate an unbiased hypothesis

Katarzyna Kulej

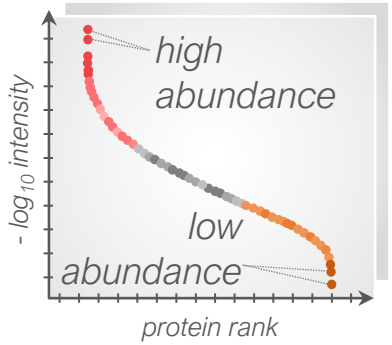
Memorial Sloan Kettering Cancer Center, New York

Why do we do it?

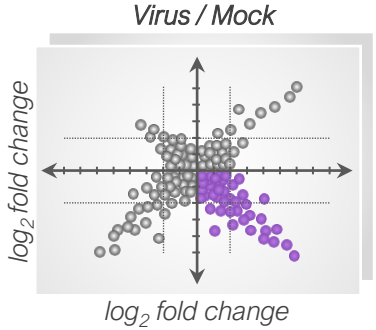
- We all produce large amounts of data, and we rarely have time to make sense of them
- The first instinct in proteomics is to calculate the fold change of our proteins between condition A vs condition B, and expect that this unravels the molecular mechanisms of our system
- While there is nothing wrong in doing so, we forget that there are many more perspectives we can observe our data from. These can help us decipher unexpected properties of our sample
- We will initially discuss these perspectives, and then the O'Donovan lab will present a specific application in metabolomics

Quantitative Dimensions in Proteomics

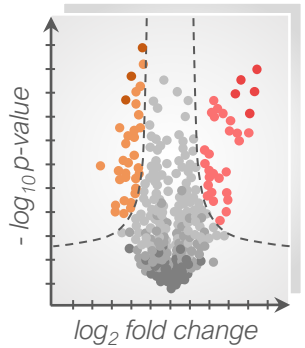
protein abundance



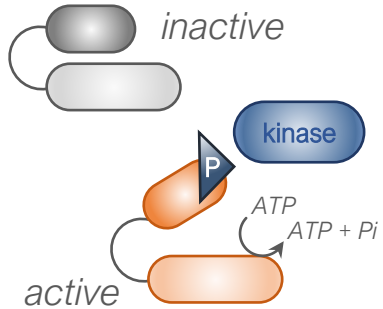
protein fold change



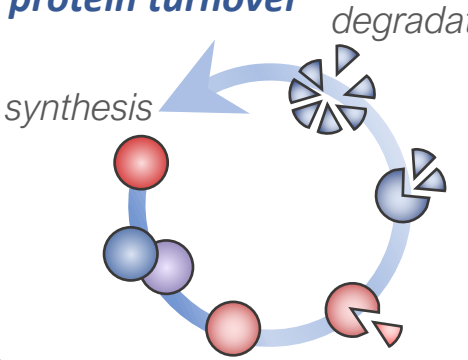
confidence (p-value)



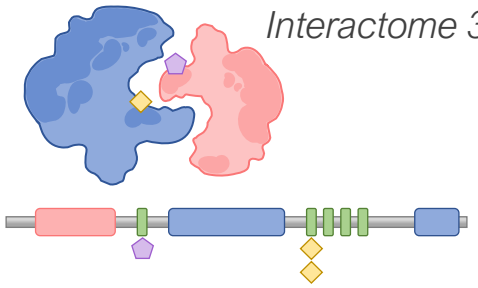
protein PTMs



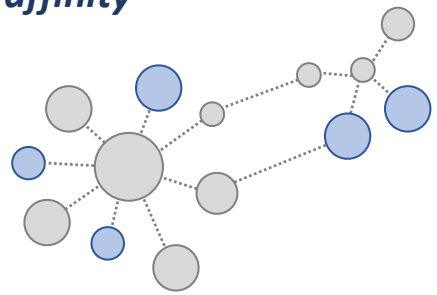
protein turnover



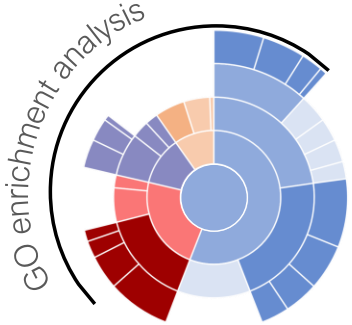
protein folding



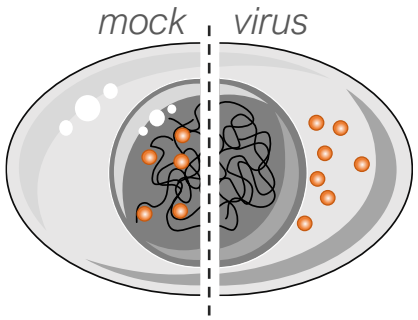
protein-protein affinity



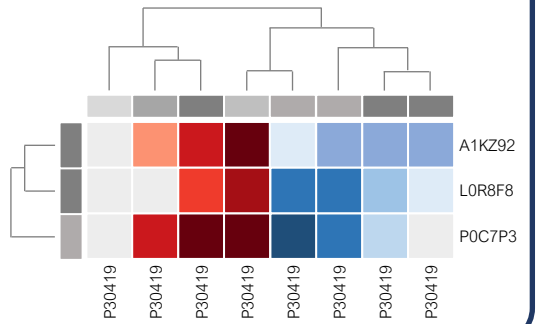
protein activity (function)

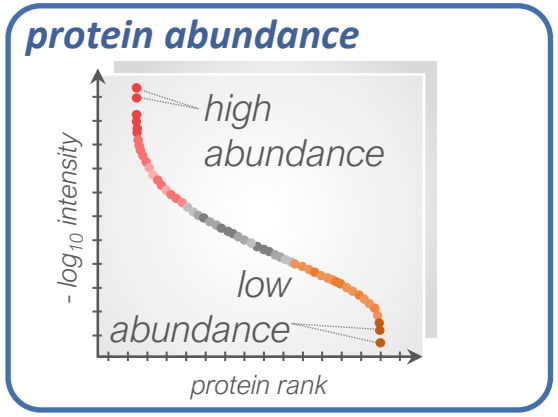


protein translocation



protein co-regulation



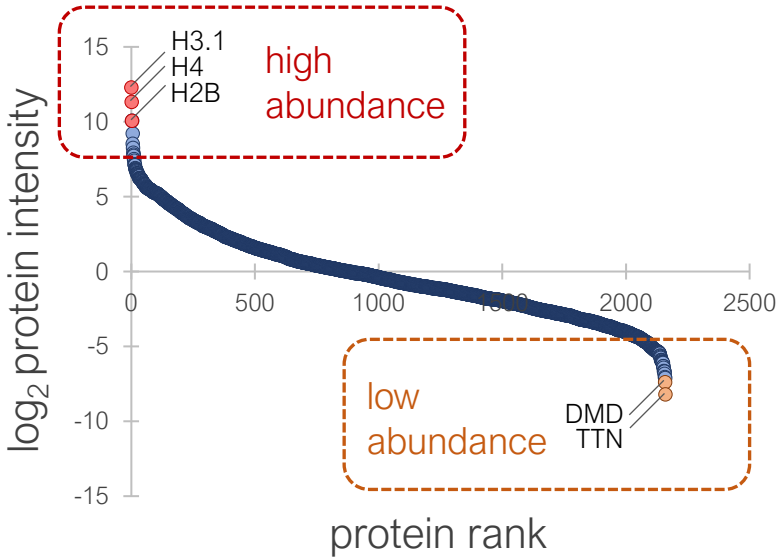


Protein abundance (what it means and how to visualize it)

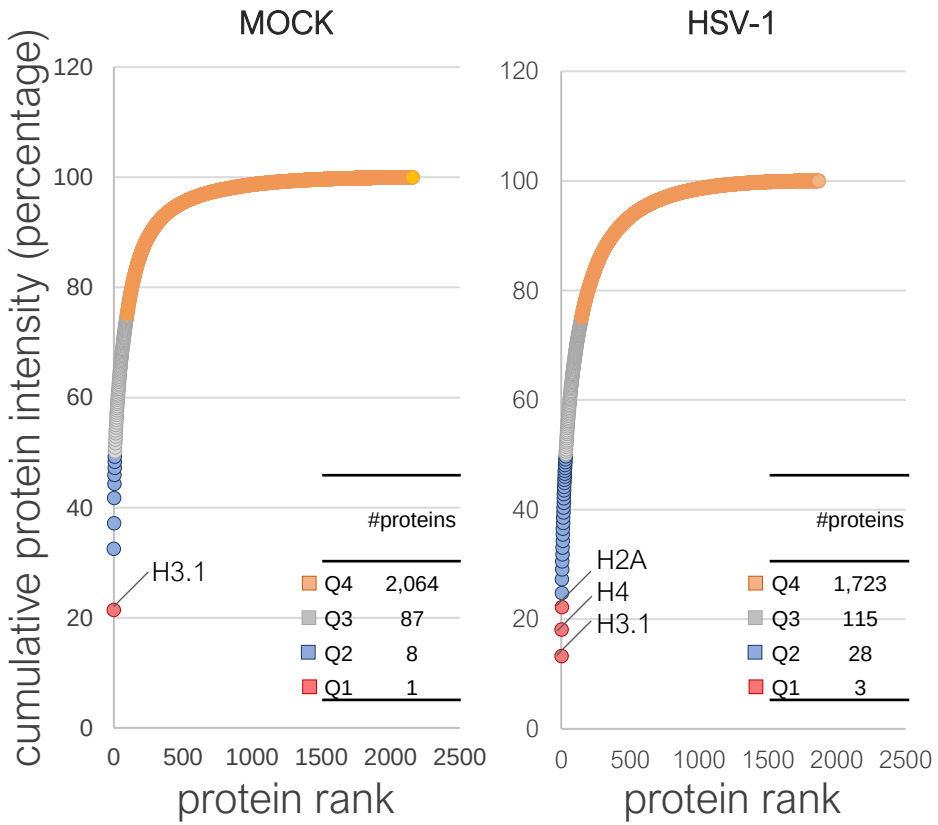
- Dynamic range issues
- Abundant proteins might be better biomarkers
- Abundance plots can be used to differentiate sample types (e.g., highly specialized cells have very few very abundant proteins, while stem cells have lots of medium abundant proteins)

Protein rank plot from highest to lowest abundant protein, illustrating the dynamic range

Host nascent DNA-bound proteome (MOCK)

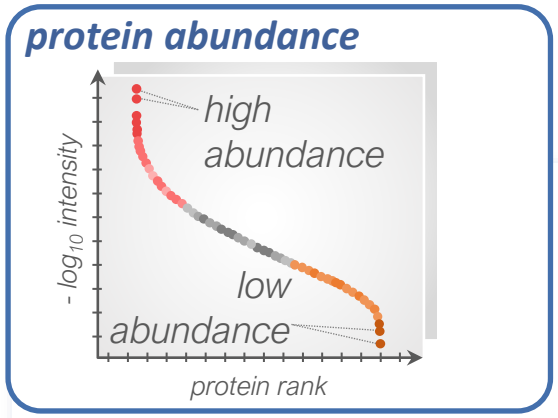


Isolation of Proteins On Nascent DNA (iPOND)

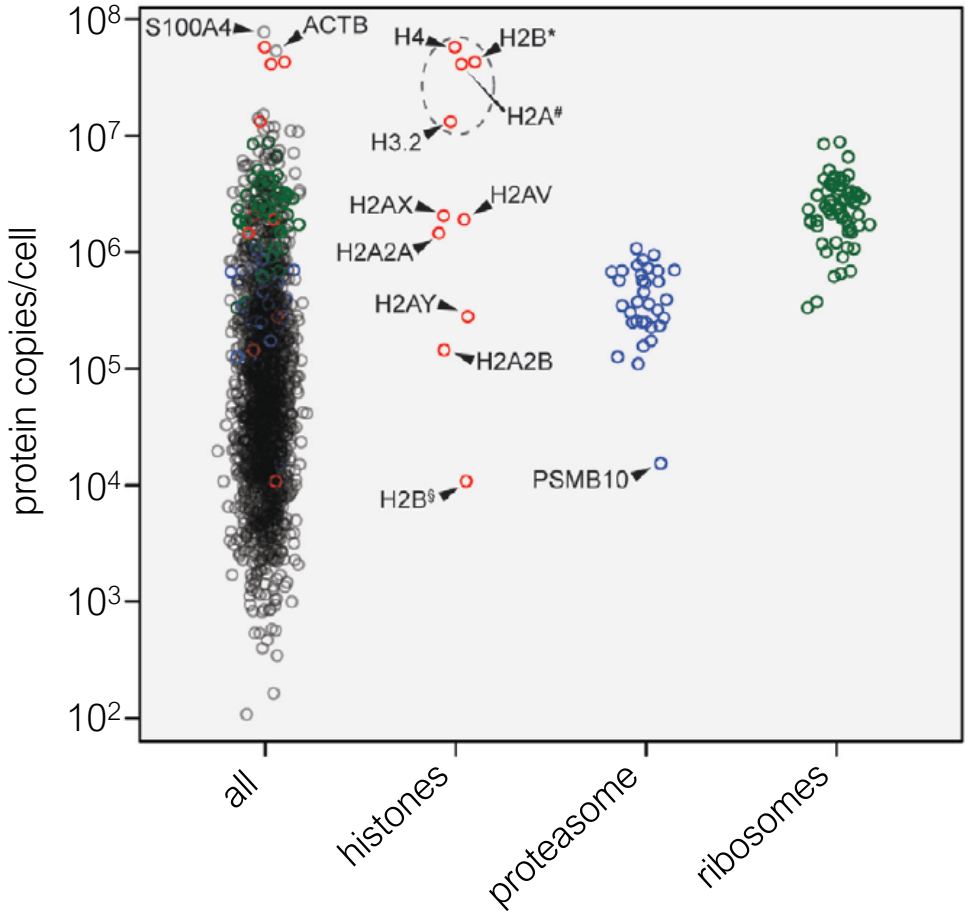


Reyes E.D. et al. (2017), Mol Cell Proteomics

Protein absolute intensity



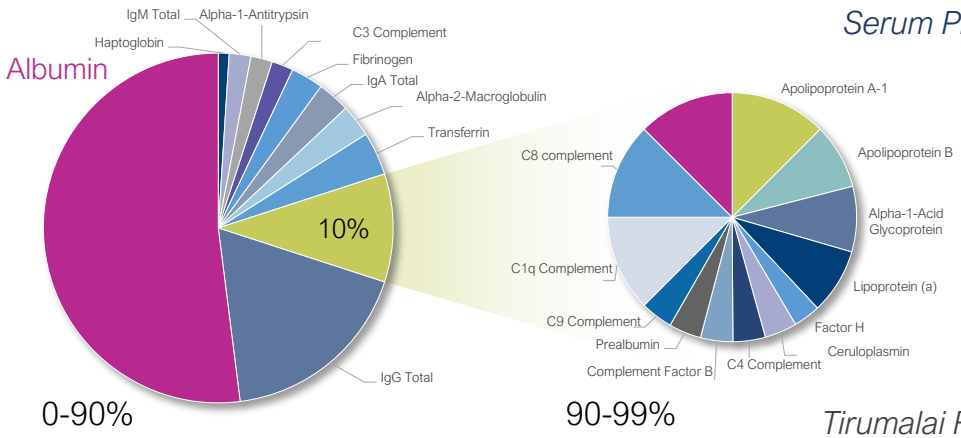
Total mouse proteome



from MaxQuant summer school, Selbach M.

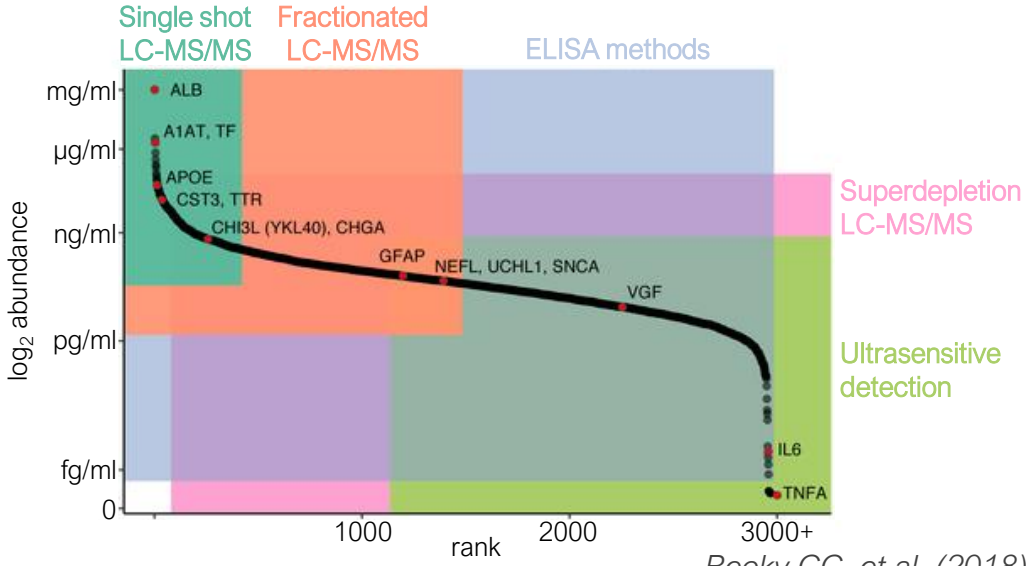
Biofluid proteome

Absolute Quantification of Serum Proteome



Tirumalai RS,. et al.,(2003), MCP

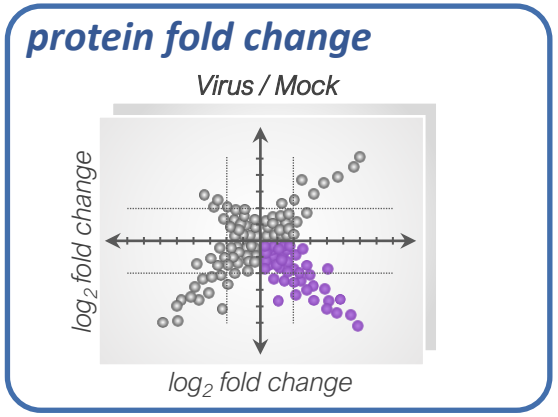
Cerebrospinal fluid (CSF) proteins ranked according to their abundance measured by various techniques.

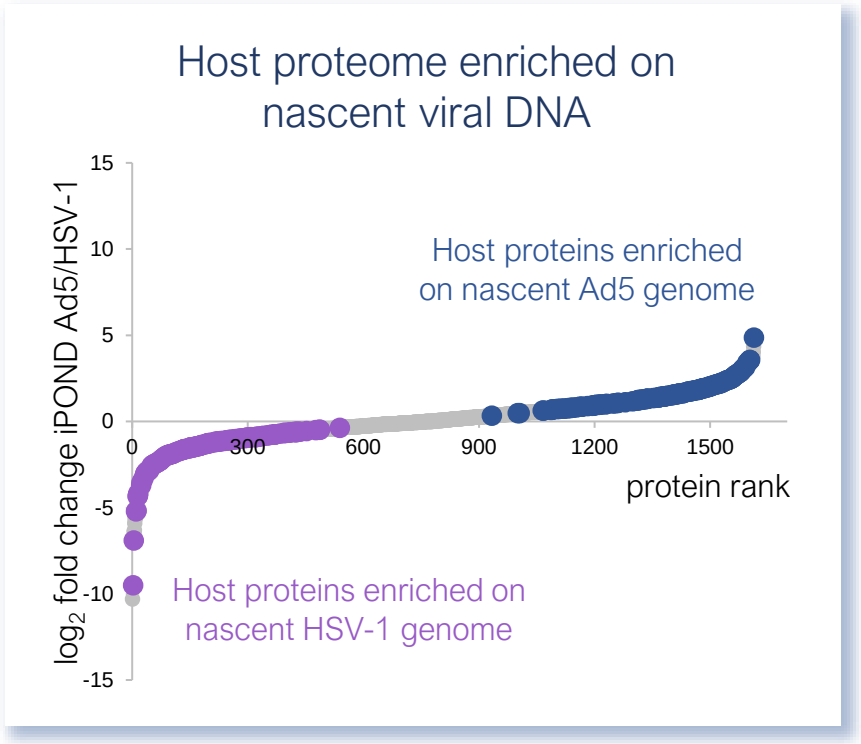


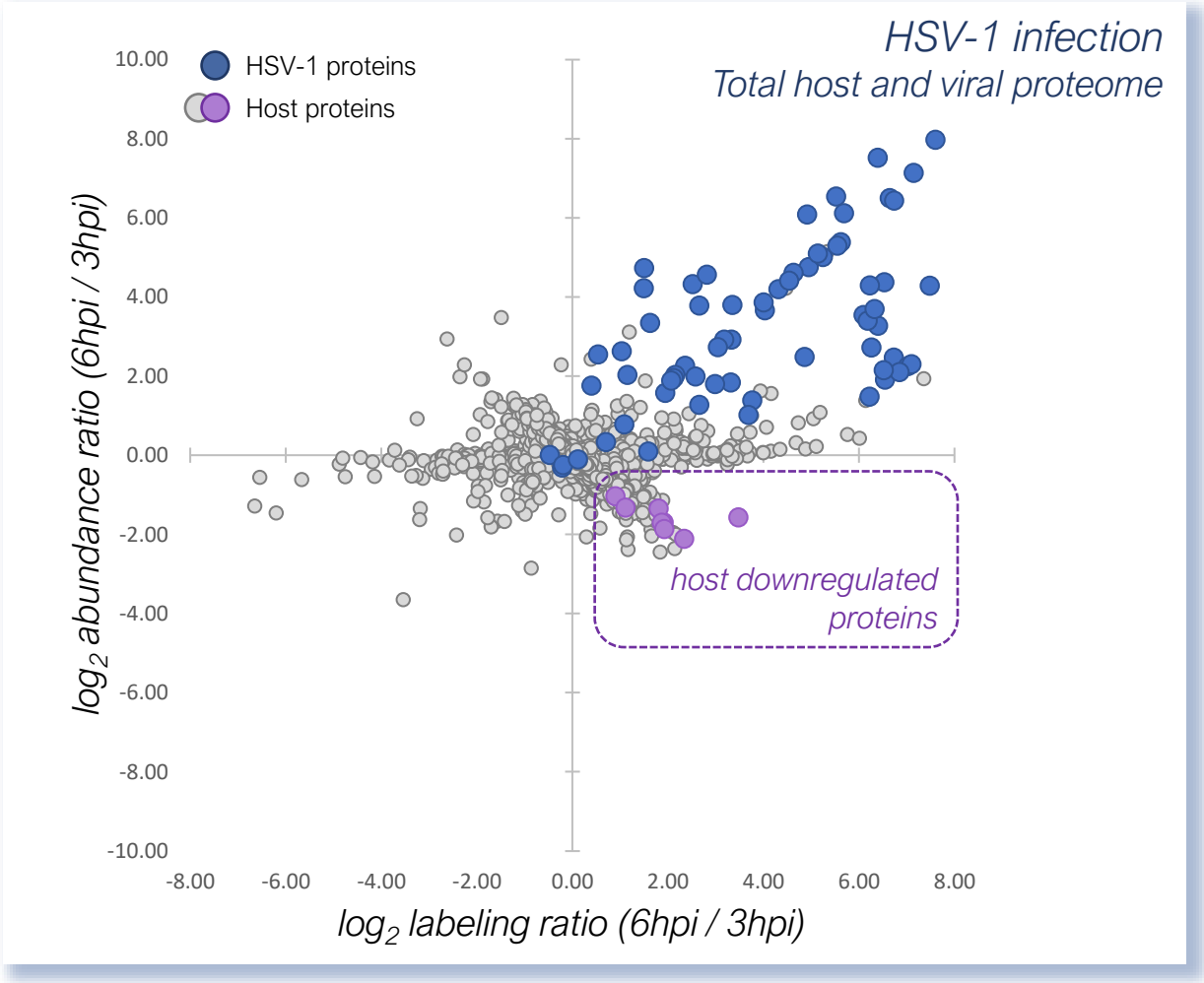
Becky CC. et al.,(2018), Proteomes

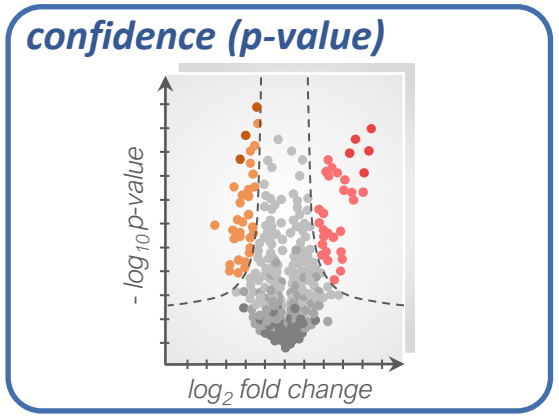
Protein fold change

- It is the most intuitive dimension
- To treat carefully: without p-value it might be meaningless, and it does not provide a perspective of how much protein there is









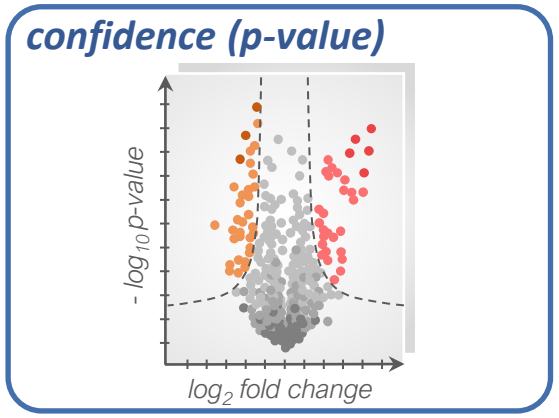
P-value and its meaning

To consult the statistician after an experiment is finished is often merely to ask him to conduct a postmortem examination.
He can perhaps say what the experiment died of.
- R.A. Fisher

The first principle is that you must not fool yourself, and you are the easiest person to fool.
- Richard P. Feynman

- Remember, a P-value is not a measure of how right you are or how important a difference is. Instead, think of it as a measure of surprise.
- If you assume your medication is ineffective and there is no reason other than luck for the two groups to differ, then the smaller the p-value, the more surprising and lucky your results are – or your assumption is wrong, and the medication truly works.

Reinhart A., Statistics done wrong (2015).



P-value and its meaning

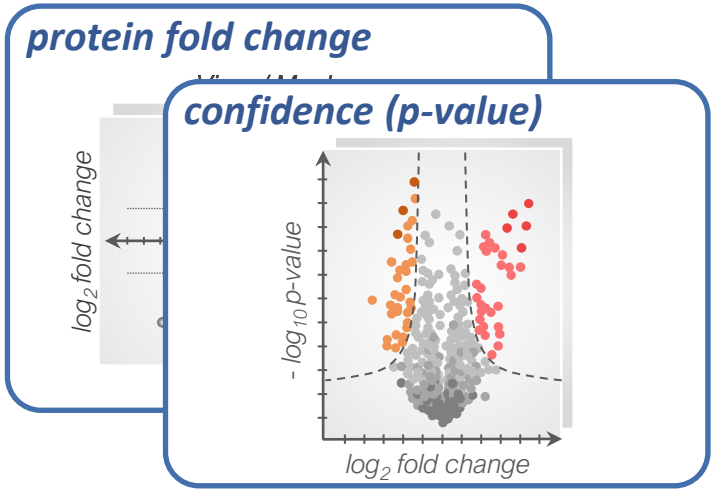
To consult the statistician after an experiment is finished is often merely to ask him to conduct a postmortem examination.
He can perhaps say what the experiment died of.
- R.A. Fisher

The first principle is that you must not fool yourself, and you are the easiest person to fool.
- Richard P. Feynman

- The P-value reports a probabilistic significance, not a biological one.
- In other words, statistical significance does not mean your result has any practical significance.
- The choice of $p < 0.05$ as significant is not because of any special logical or statistical reasons but it has become a scientific convention through decades of common use.

Krzywinski M. & Altman N., et al. (2013), Nature Methods; Reinhart A., Statistics done wrong (2015).

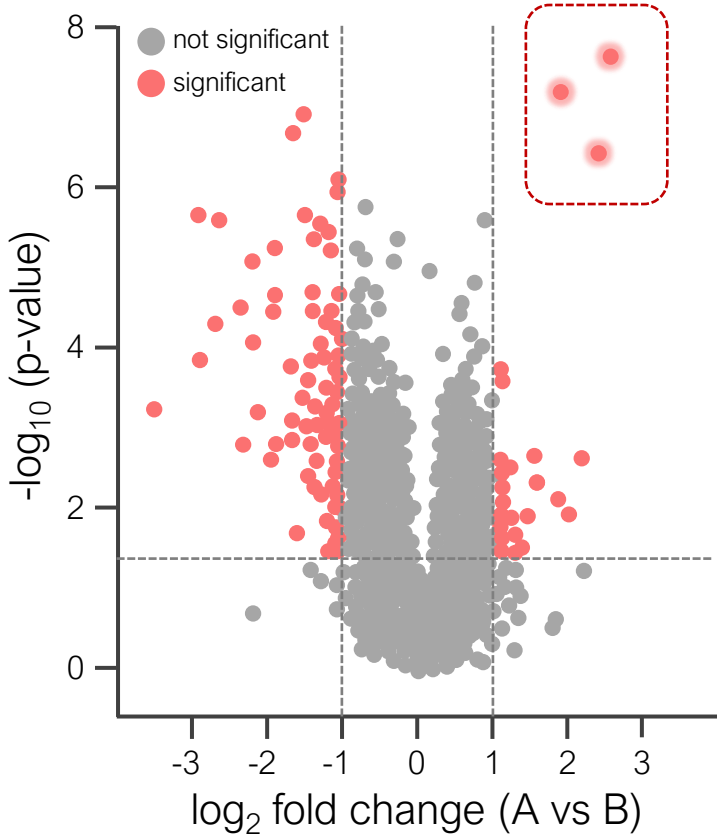
First (obvious) examples of data analysis by combining quantitative dimensions



Fold change AND p-value

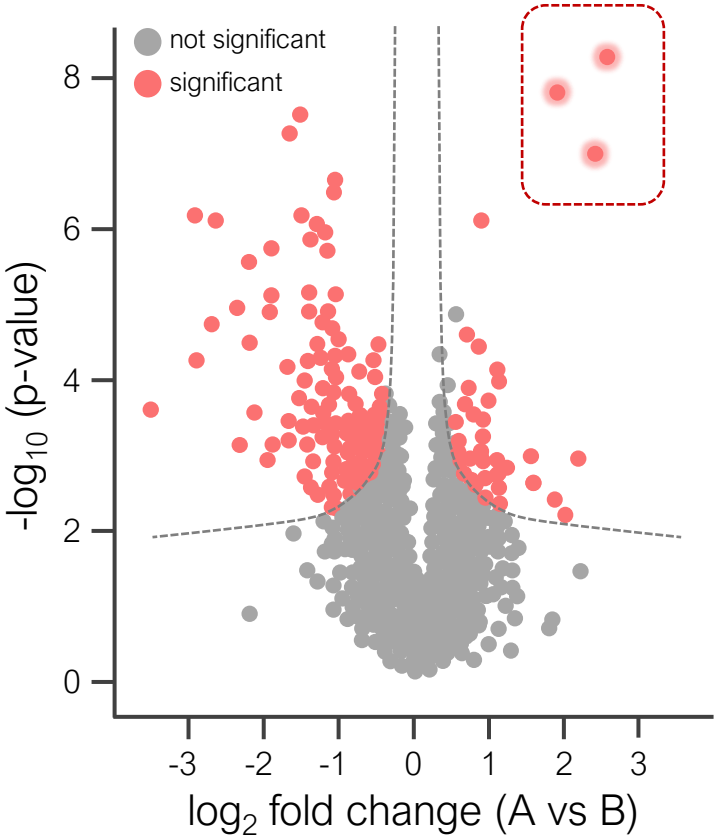
Volcano plots showing results of comparisons between two conditions (A vs B).

The thresholds for calling a protein <u>differentially abundant</u> can be defined by one of two methods:



(i) square cutoffs for p-value and fold-change

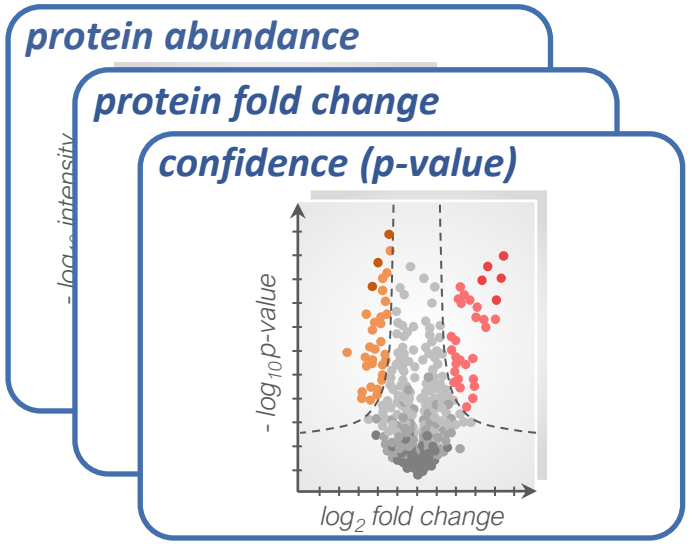
- multiple hypothesis testing corrected p-value (or FDR)
- fold-change cutoff is set with regard to the experimental power



(ii) non-linear volcano lines

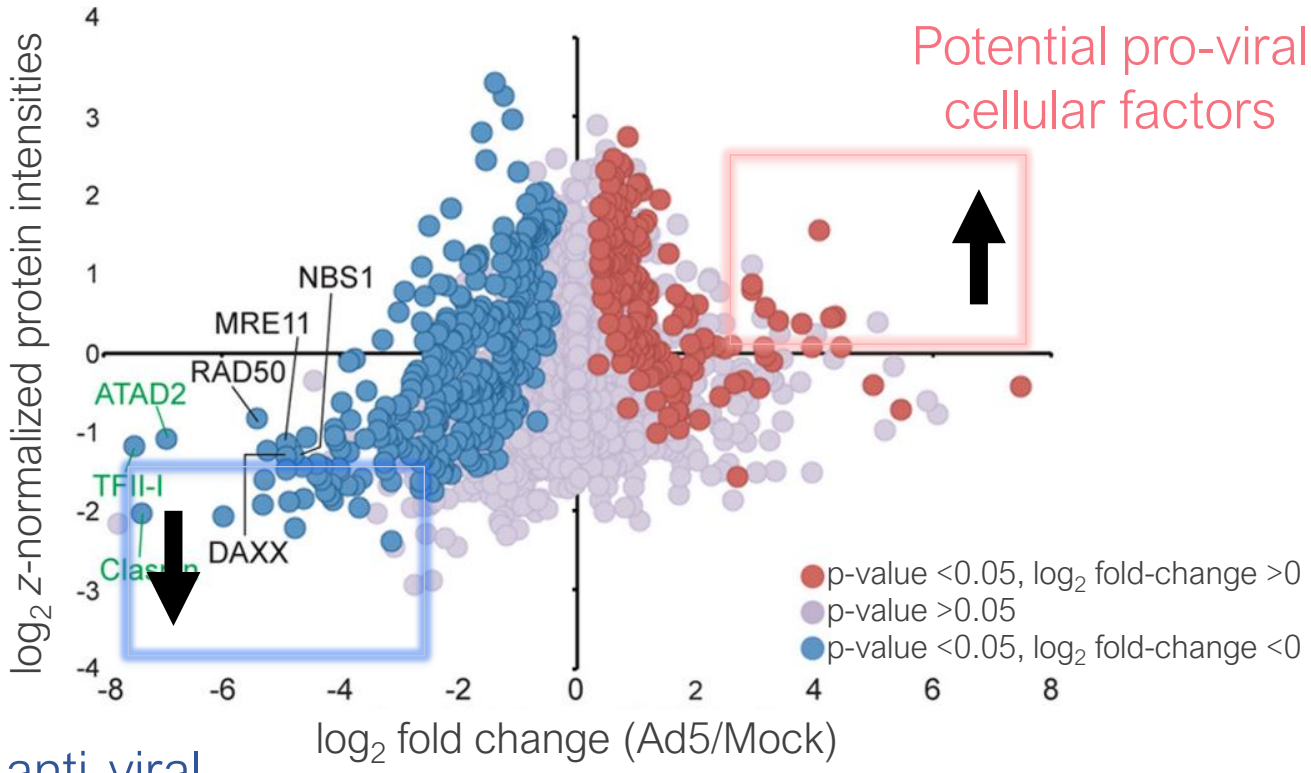
- FDR
- s0 parameter is set instead of a specific fold-change cutoff

Schessner J.P., et al. (2022) Proteomics.



Abundance AND fold change AND p-value

Infection of host cells (Mock) with Adenovirus type-5 (Ad5)



Potential anti-viral cellular factors

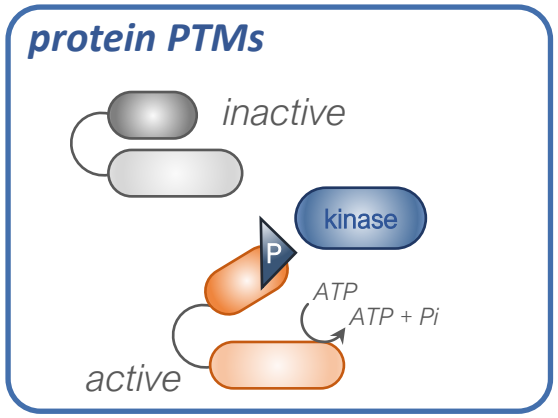
Use the absolute/relative intensity!

- You can find biomarkers
- You can prioritize candidates
- You can interpret protein hindrance

Reyes E.D. et al. (2017), Mol Cell Proteomics

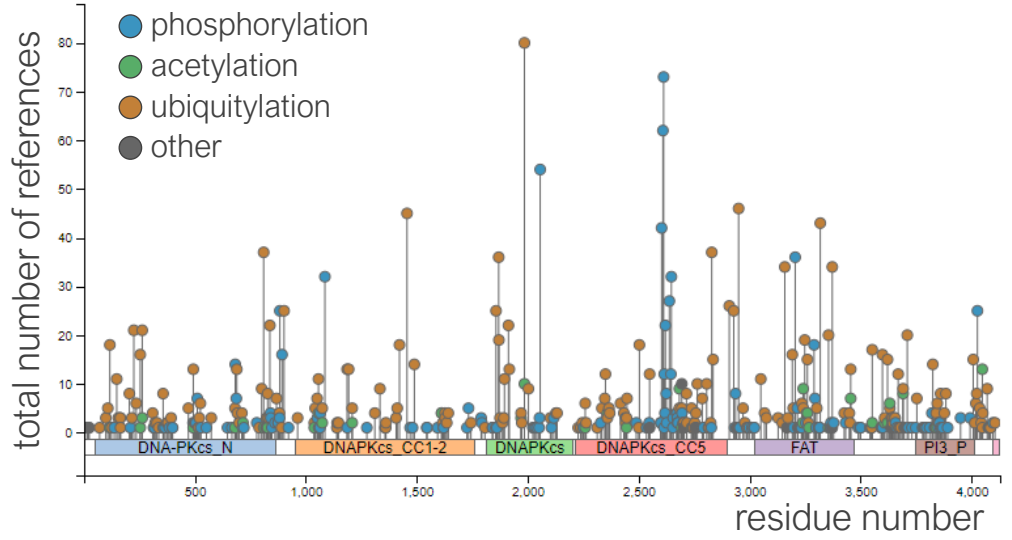
Less obvious quantitative dimensions

Protein post-translational modifications (PTMs)

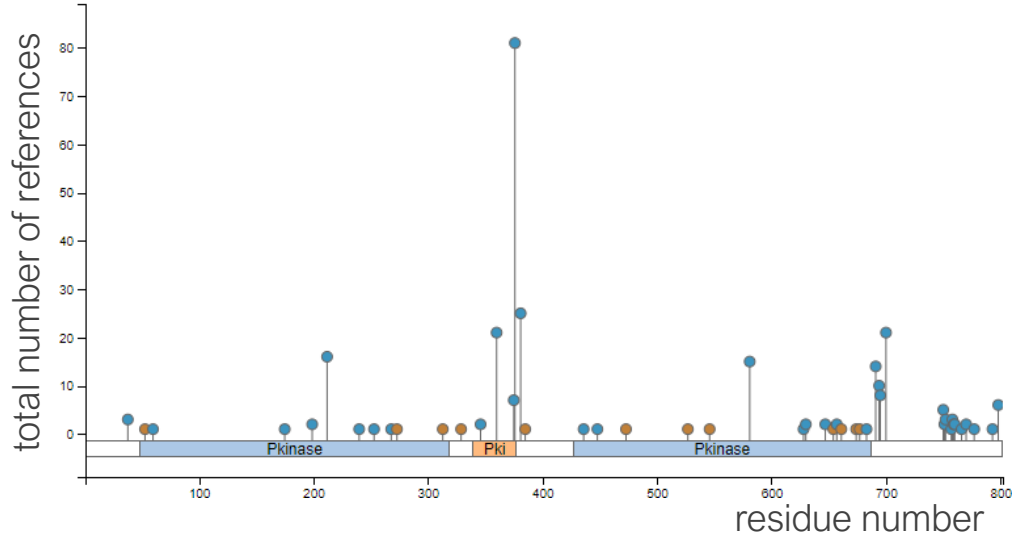


- Protein PTMs have their own abundance, independent from protein abundance
- PTM stoichiometry can be used to assess activity of the enzymes that catalyze them, and thus predict biological targets

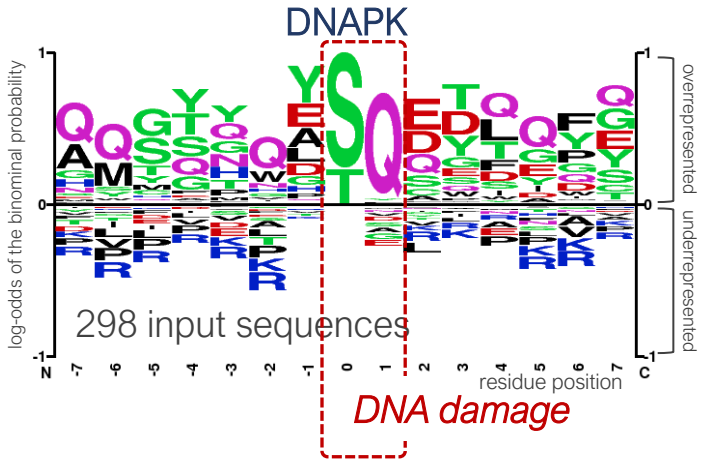
DNAPK - DNA-dependent protein kinase catalytic subunit

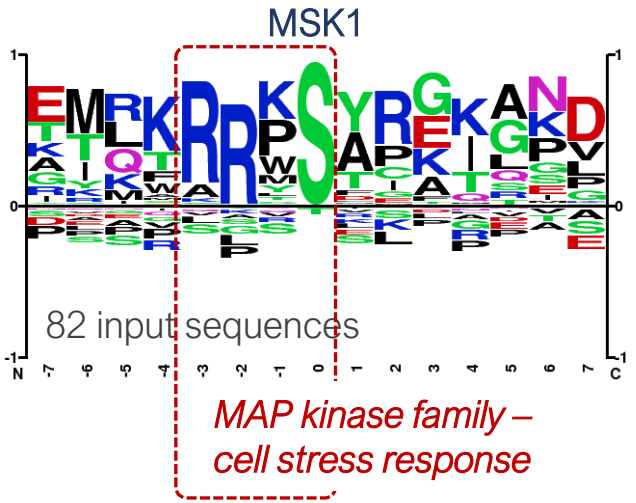


MSK1 - Ribosomal protein S6 kinase alpha-5



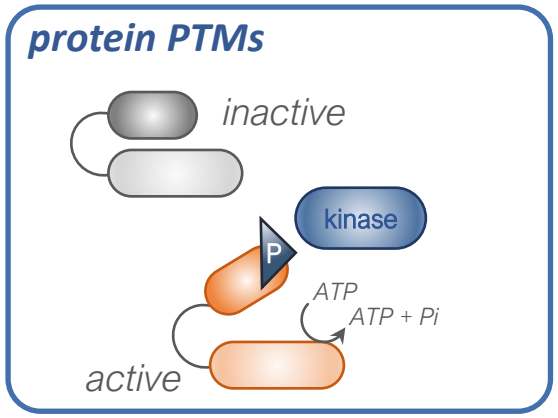
Kinase substrate motif



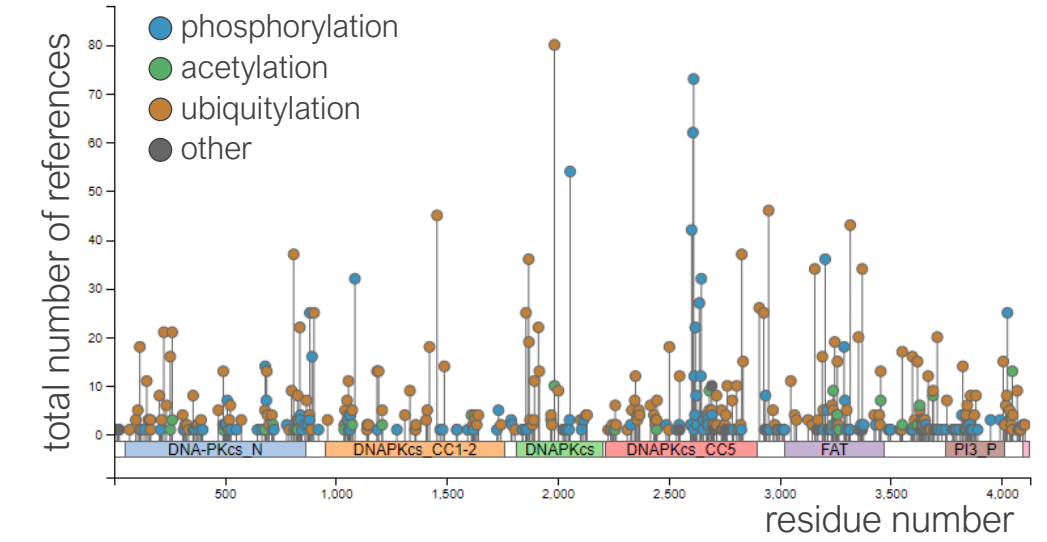


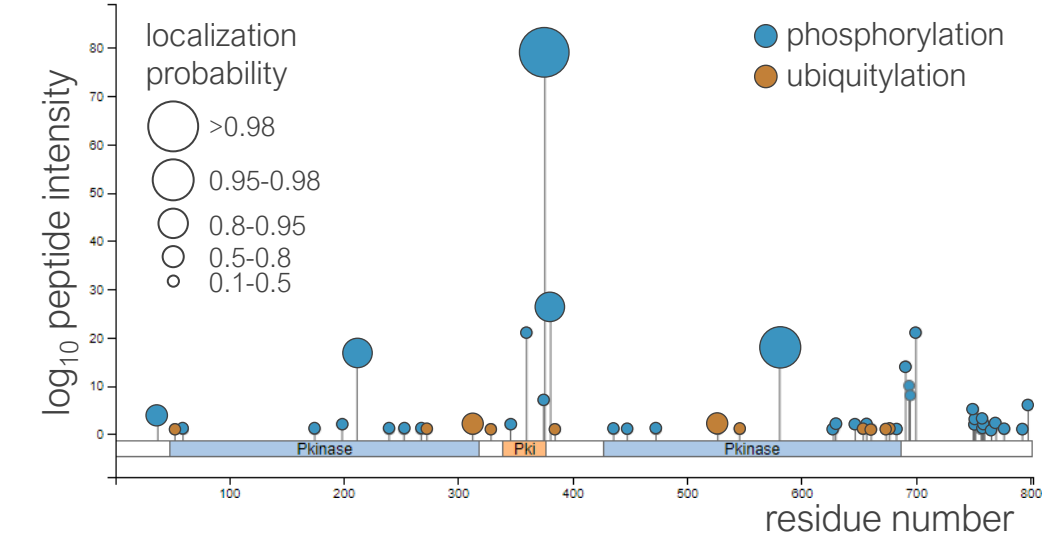
https://www.phosphosite.org/homeAction

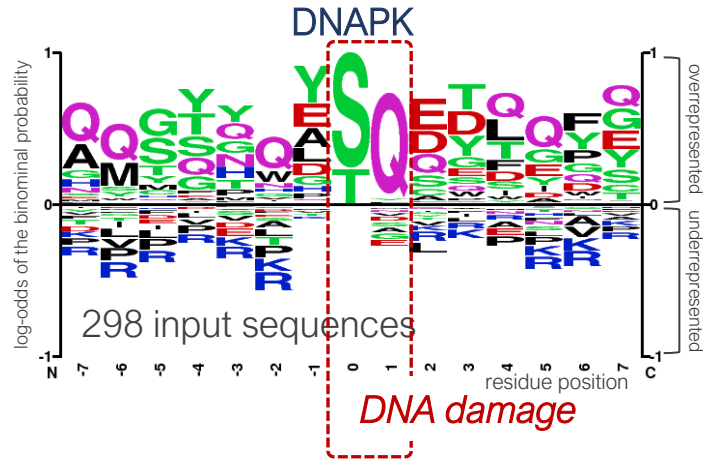
Protein post-translational modifications (PTMs)

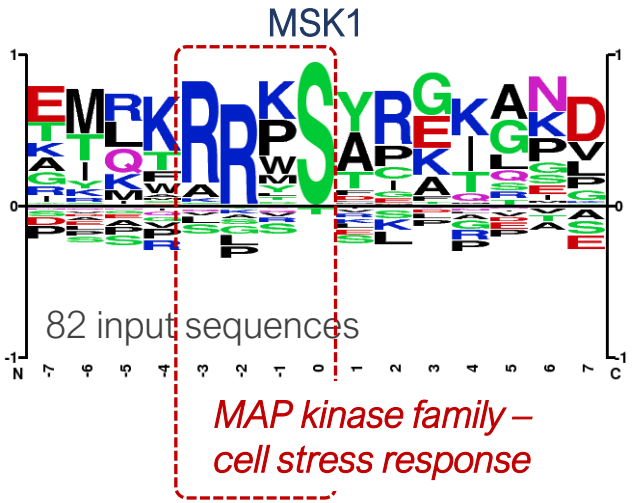


Confidence in PTM localization is also quantitative, and it can help to prioritize targets



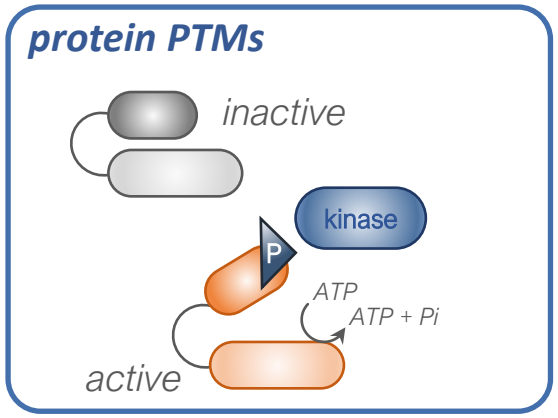






https://www.phosphosite.org/homeAction

Synaptosomes kinome - iTOL



Interactive Tree Of Life

Information obtained from previously published data

PostSynaptic Density (PSD) cellular structure

Active Zone (CAZ) of the neurotransmitter

Protein domains

Protein phosphosite regulation in different conditions

PSD/CAZ localization
- Collins M.O. *et al*. 2006
- Bayes A. *et al*. 2011; 2012
- Distler U. *et al*. 2014
- Boyken J. *et al*. 2013

Domains
- Phosp-site within kinase domain
- Transmembrane domiain(s)
- Signal peptide

Phosphorylation sites regulation

Site1
Site2
Site3
Site4
Site5
…

-4.5 0 4.5

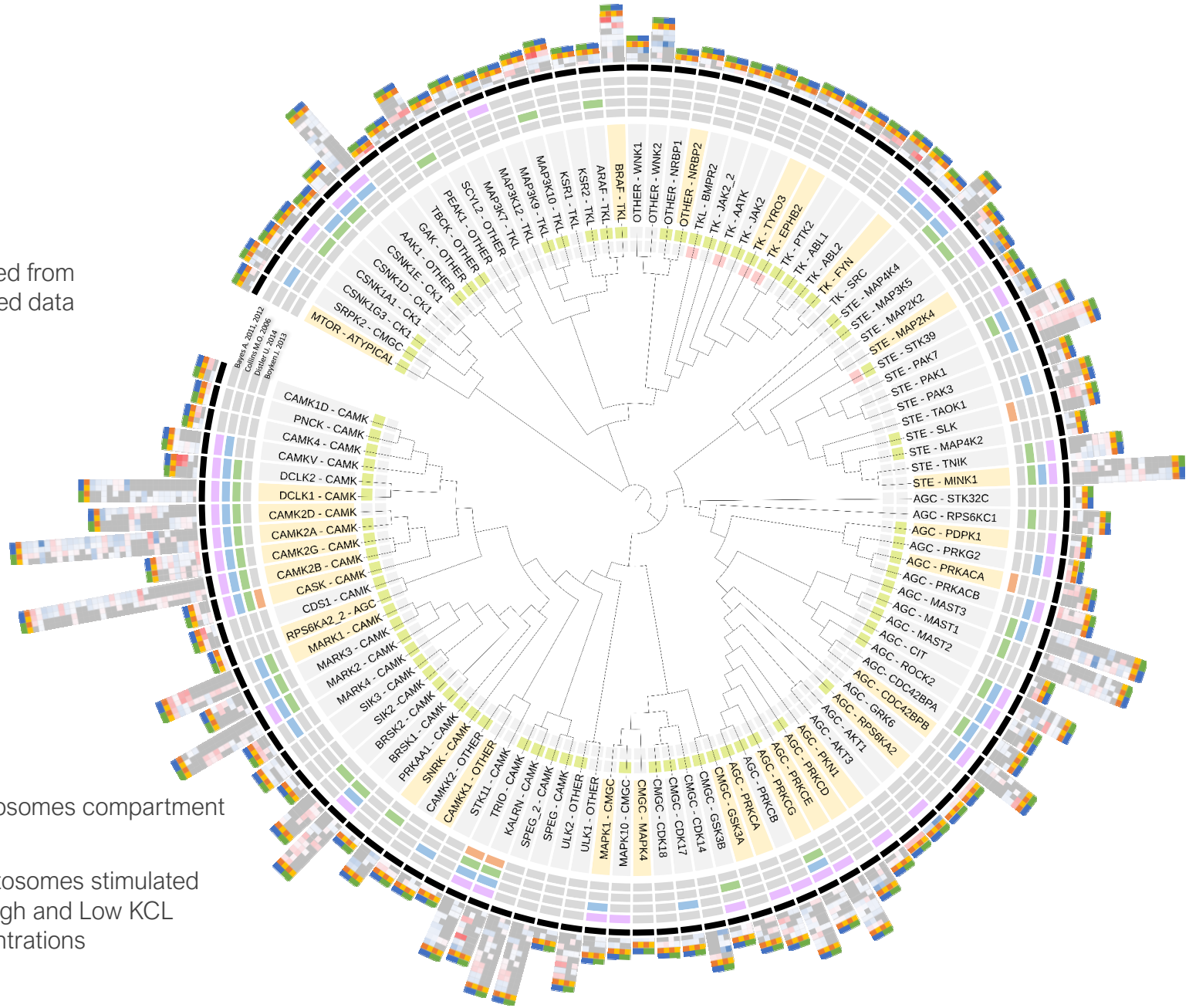
log_2 fold change

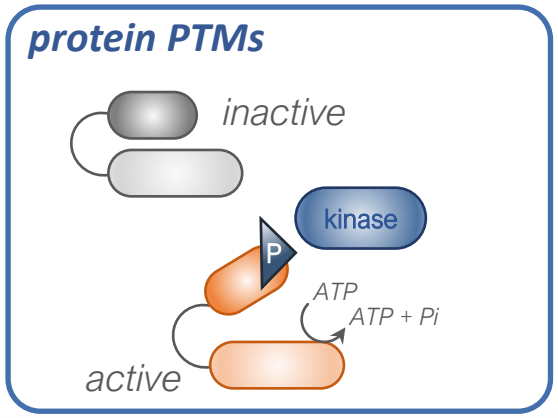
Not quantified

- Soluble fraction
- Insoluble fraction
- Low KCL stimulation
- High KCL stimulation

Synaptosomes compartment

Synaptosomes stimulated with High and Low KCL concentrations

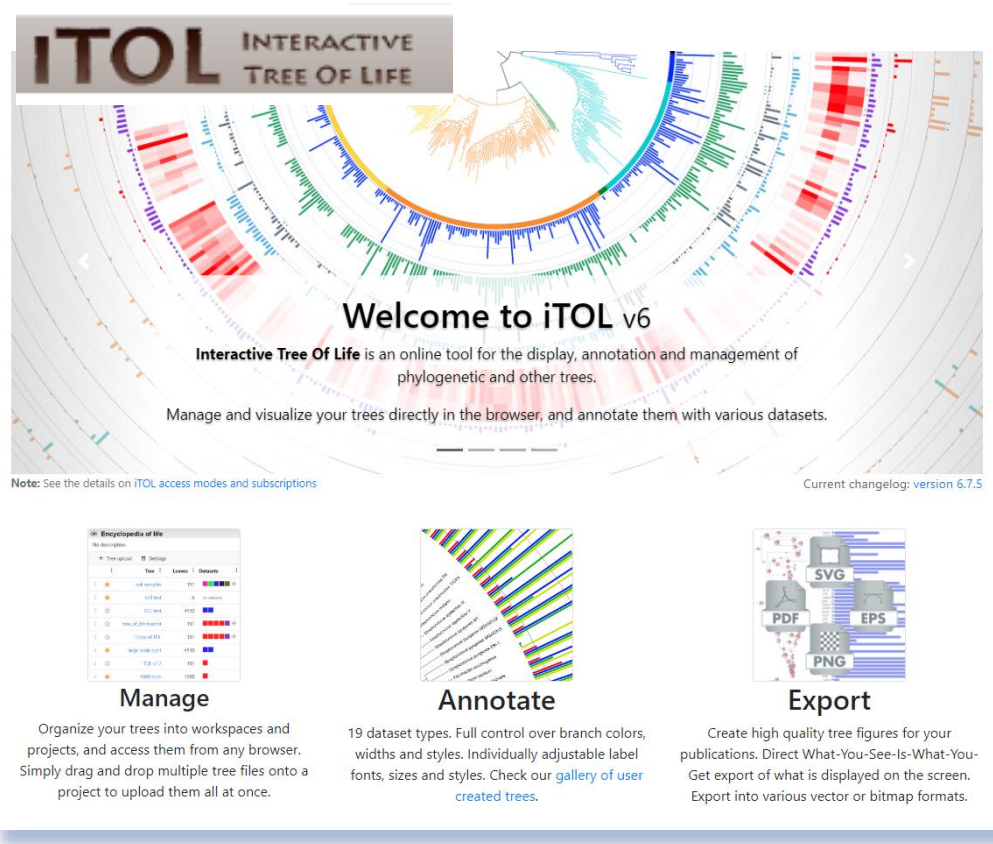


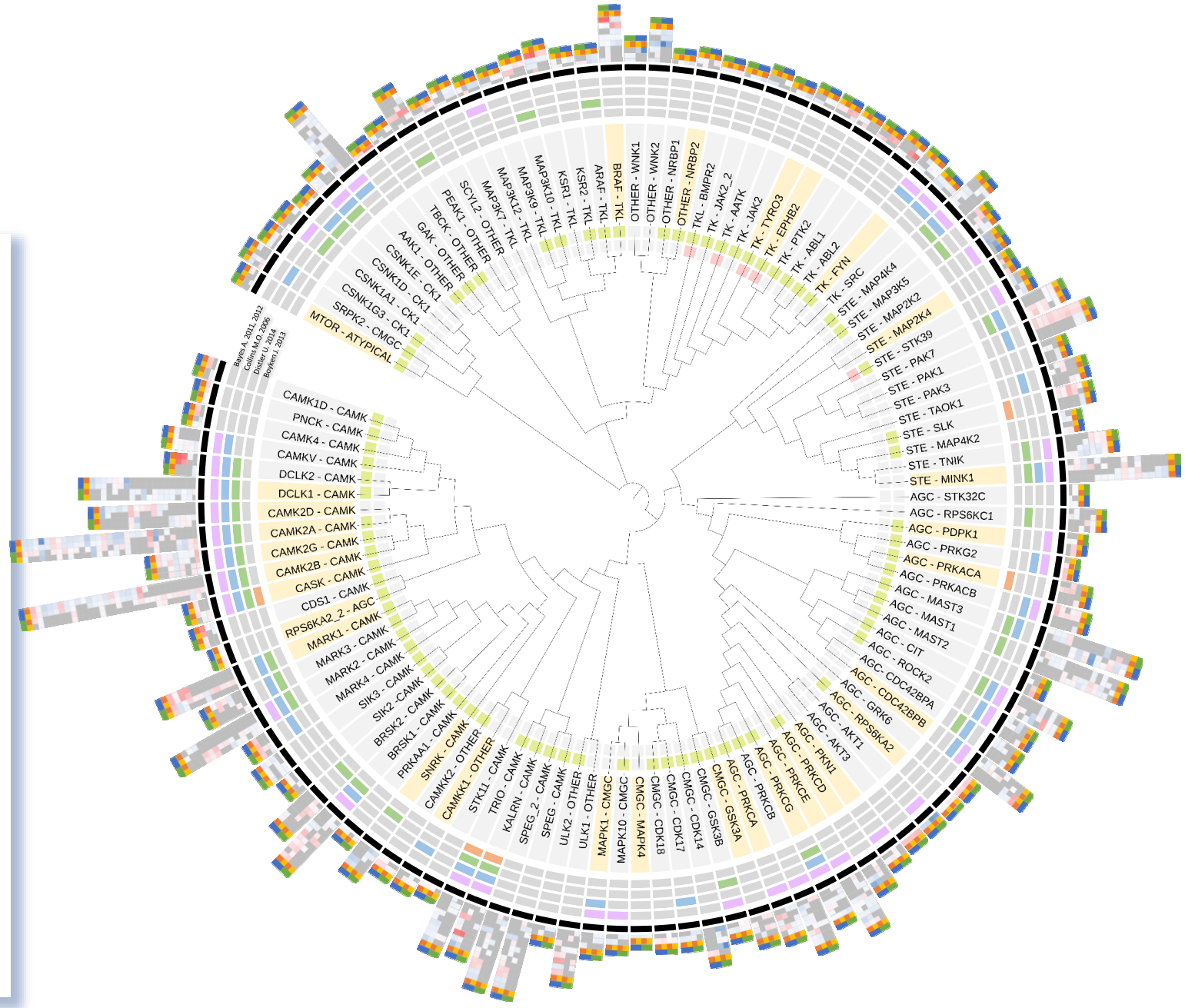


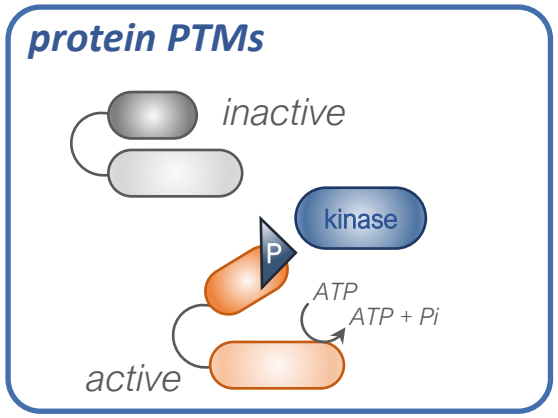
Synaptosomes kinome - iTOL

Interactive Tree Of Life

https://itol.embl.de/

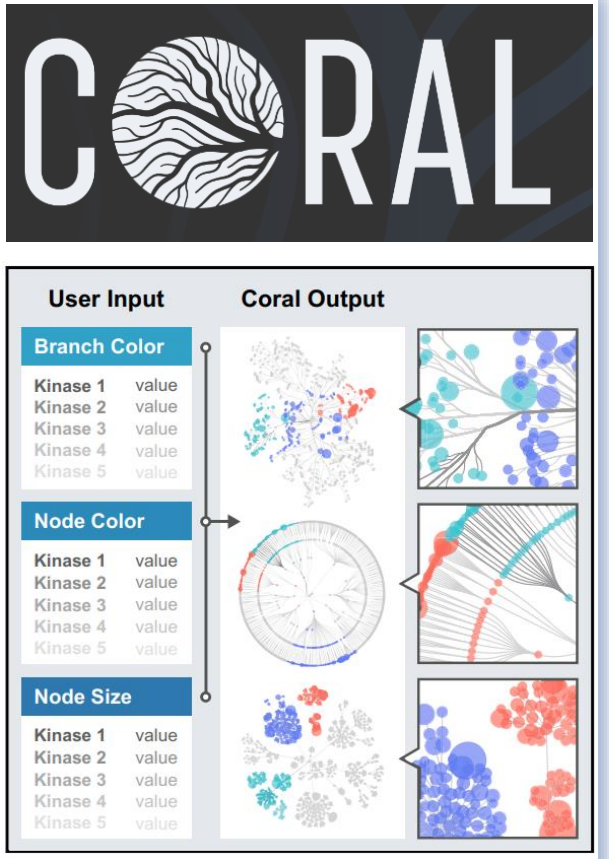




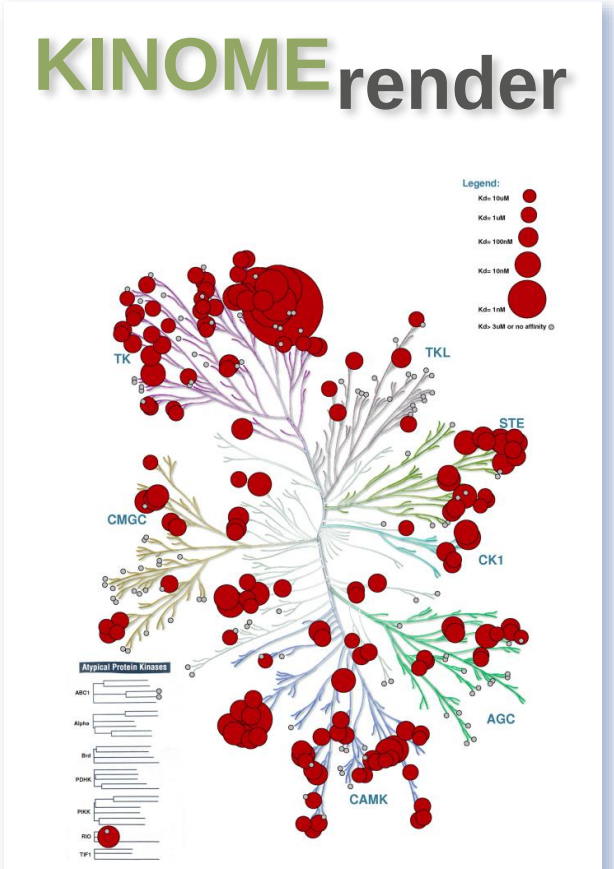


Kinome Tree

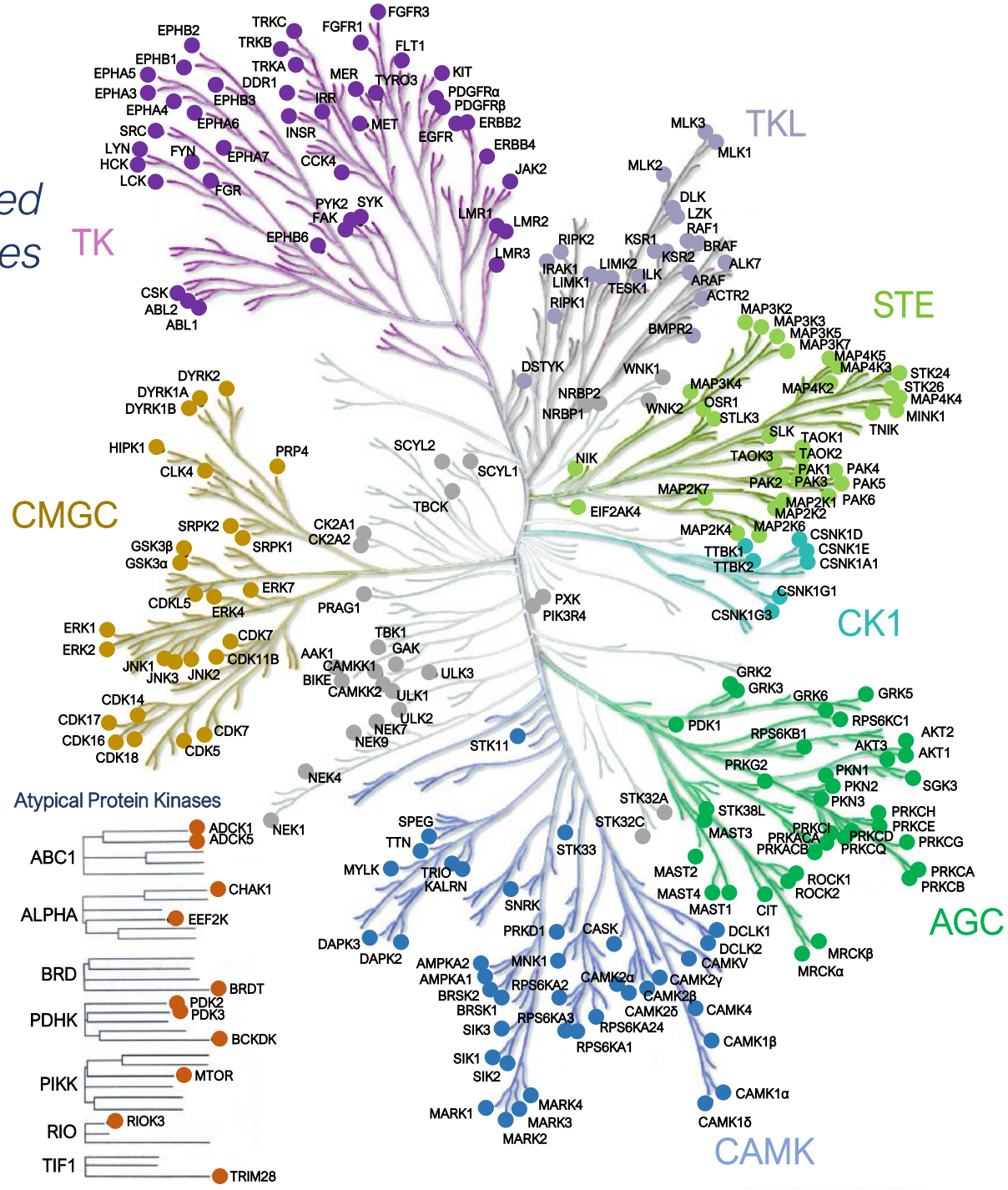
Protein kinase families enriched in synaptosomes

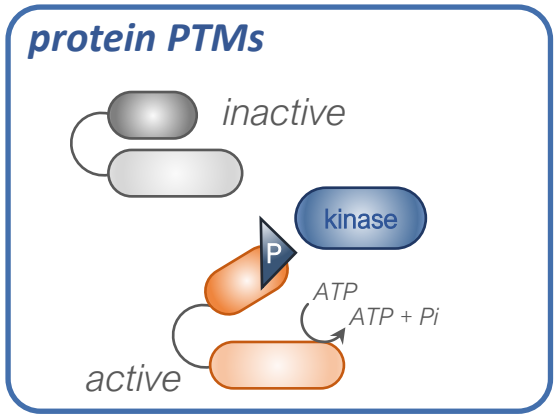


http://phanstiel-lab.med.unc.edu/CORAL/
Metz K.S. et al., (2018), Cell Systems



Najmanovich Research Group
Chartier M. et al. (2013), PeerJ





Tools for PTM assignment and quantification

Already reported PTMs

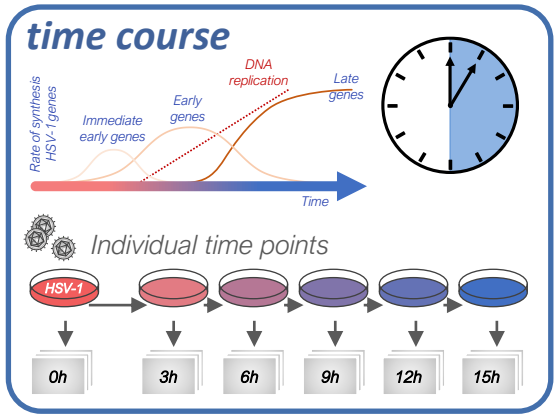
Uniprot: Experimental/putative PTMs
PhosphoELM: Curated database of validated ph-sites
PhosphoSitePlus: Experimentally reported ph-, ac- and ub-sites
OGlycBase: Experimentally reported glycosylation sites
HPRD: Experimentally reported human PTMs
dbPTM: Compendium of experimental and putative PTMs from sources above
PTMcode: Focus on PTM crosstalk
ProteomicsDB: multi-omics and multi-organism resource
HPRD: domain architecture, post-translational modifications, interaction networks and disease association for each protein in the human proteome.
PhosphoNetworks: a database for human phosphorylation networks

PTM prediction

NetPhos: Prediction of ph-sites based on the learning model
NetPhosK: Prediction of kinase-specific ph-sites
Scansite: Motifs likely to be phosphorylated by specific kinases
LysAcet: Prediction of acetylation sites
PTMfunc: Repository of functional predictions of PTMs
iGPS: prediction of site-specific kinase-substrate relations

Related kinases/transferases

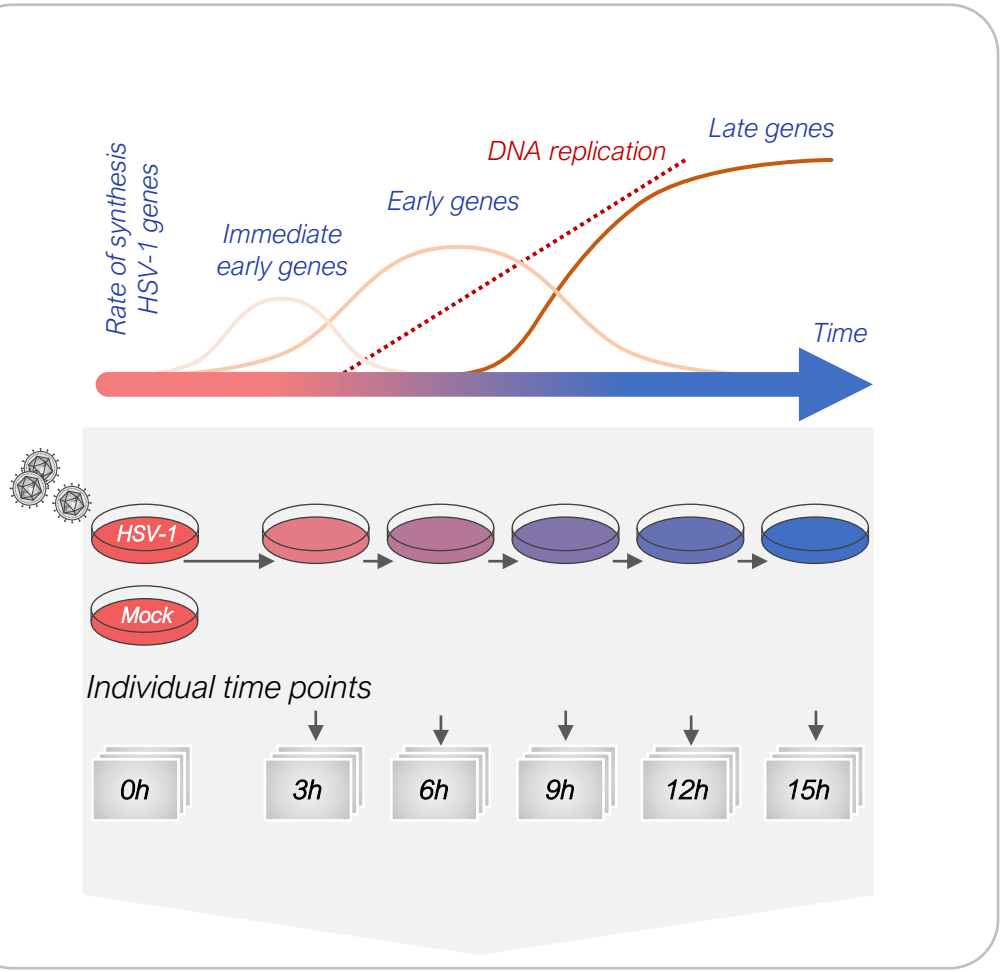
KinomeXplorer: Predict kinase-substrate interactions (NetworKIN + NetPhorest)
PKIS: Computational identification of protein kinases
PSEA: Enrichment analysis for prediction of kinases



Time series

Example 6 time points of infected cells (proteome)

HSV-1 infection

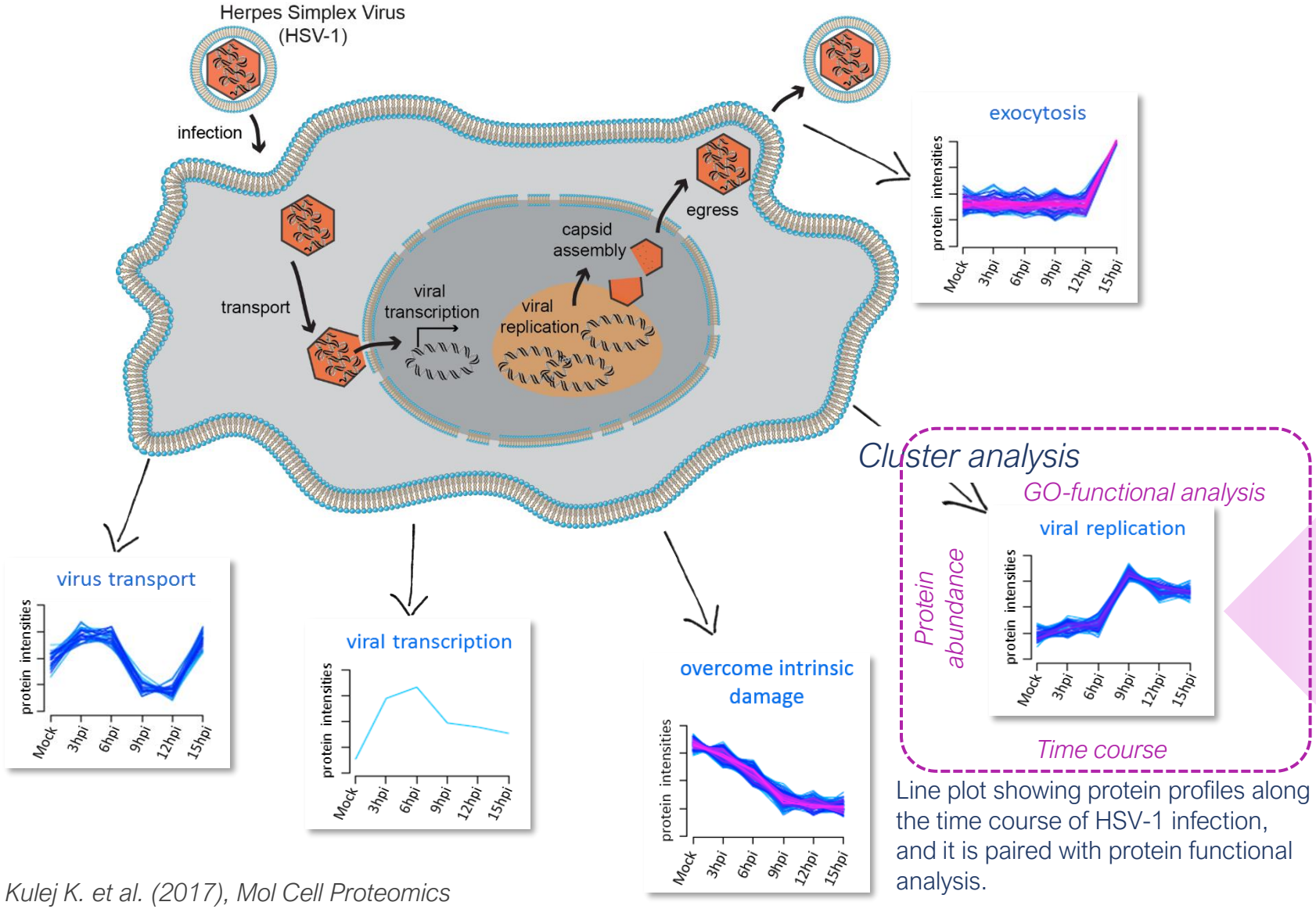


In most experiments, the proteome (and other biomolecules) is regulated in a dynamic and different manner

Distinguishing early from late responders resolves critical aspects of the biological system

Grouping trends with unsupervised clustering (e.g., fuzzy c-means)

Time course of HSV-1 infection



Cluster analysis

GO-functional analysis

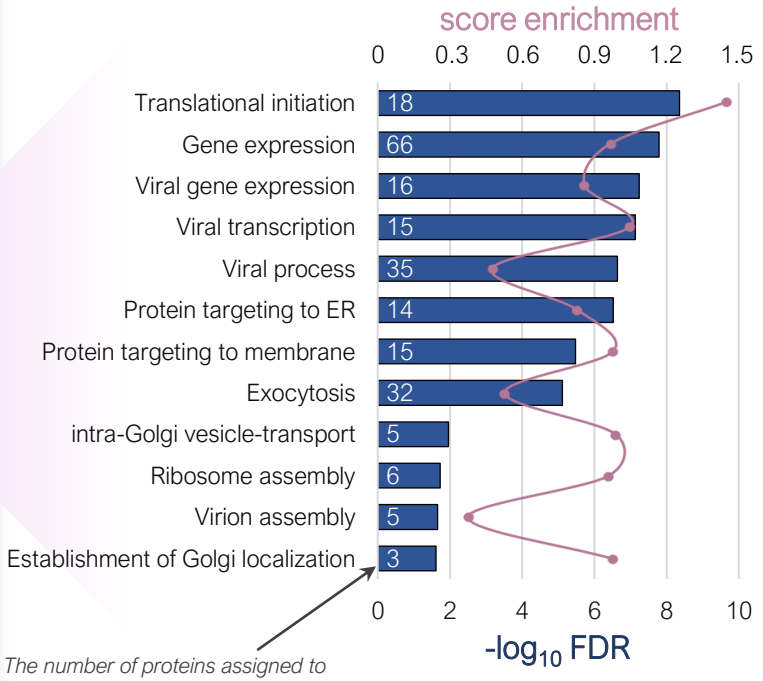
Protein abundance

Time course

Line plot showing protein profiles along the time course of HSV-1 infection, and it is paired with protein functional analysis.

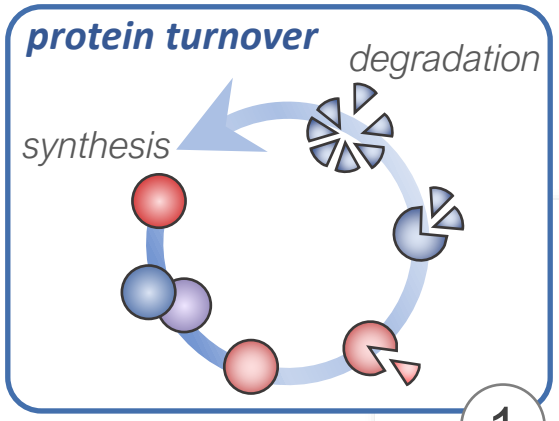
Time-resolving protein dynamic changes helps group different stages of regulation

Cluster Gene Ontology (GO) Analysis



The number of proteins assigned to classified functions is indicated by the number (white) in a bar.

Kulej K. et al. (2017), Mol Cell Proteomics



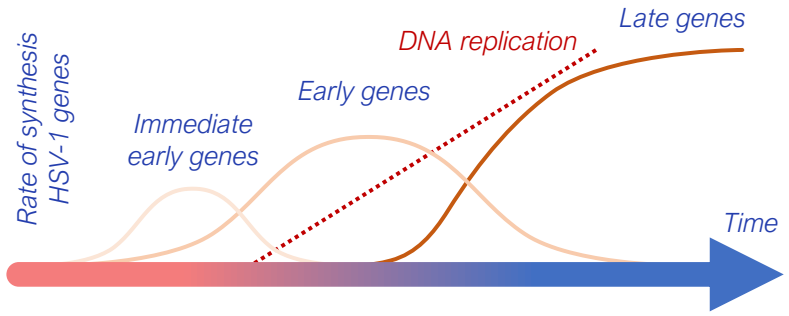
Protein synthesis rate (turnover)

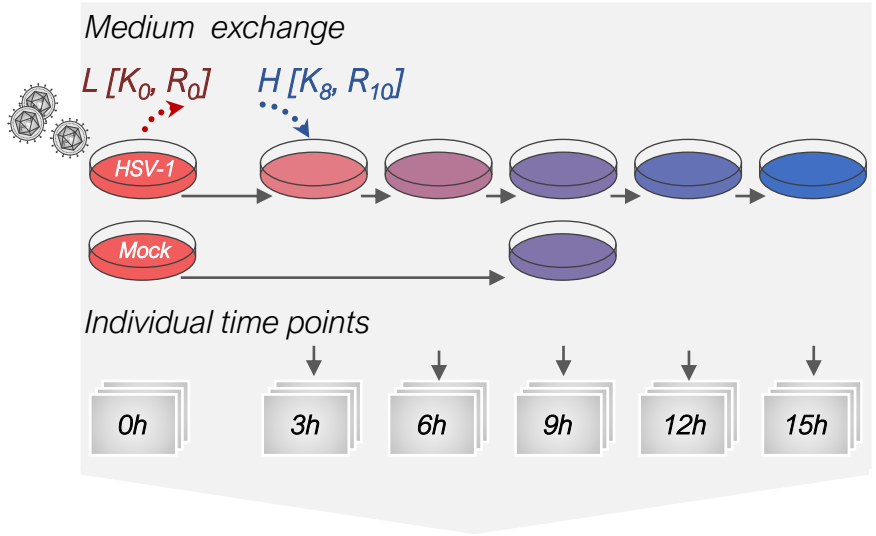
pSILAC of HSV-1 infected cells

Whole-cell proteome / phosphoproteome

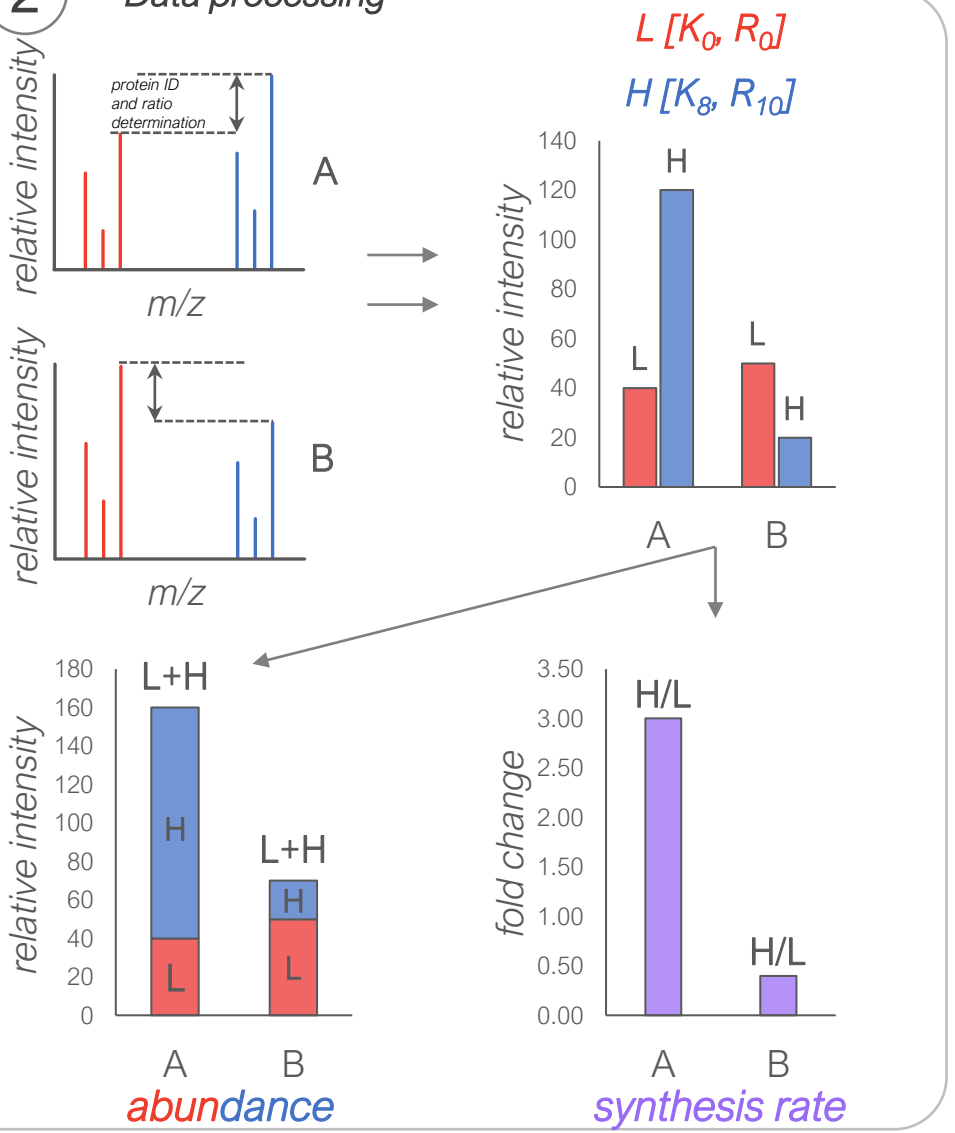
1 pSILAC labeling

pulsed stable isotope labeling by amino acids in cell culture

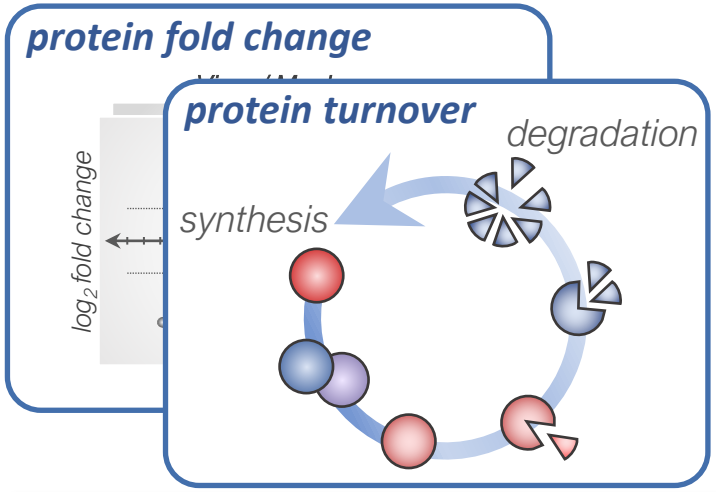


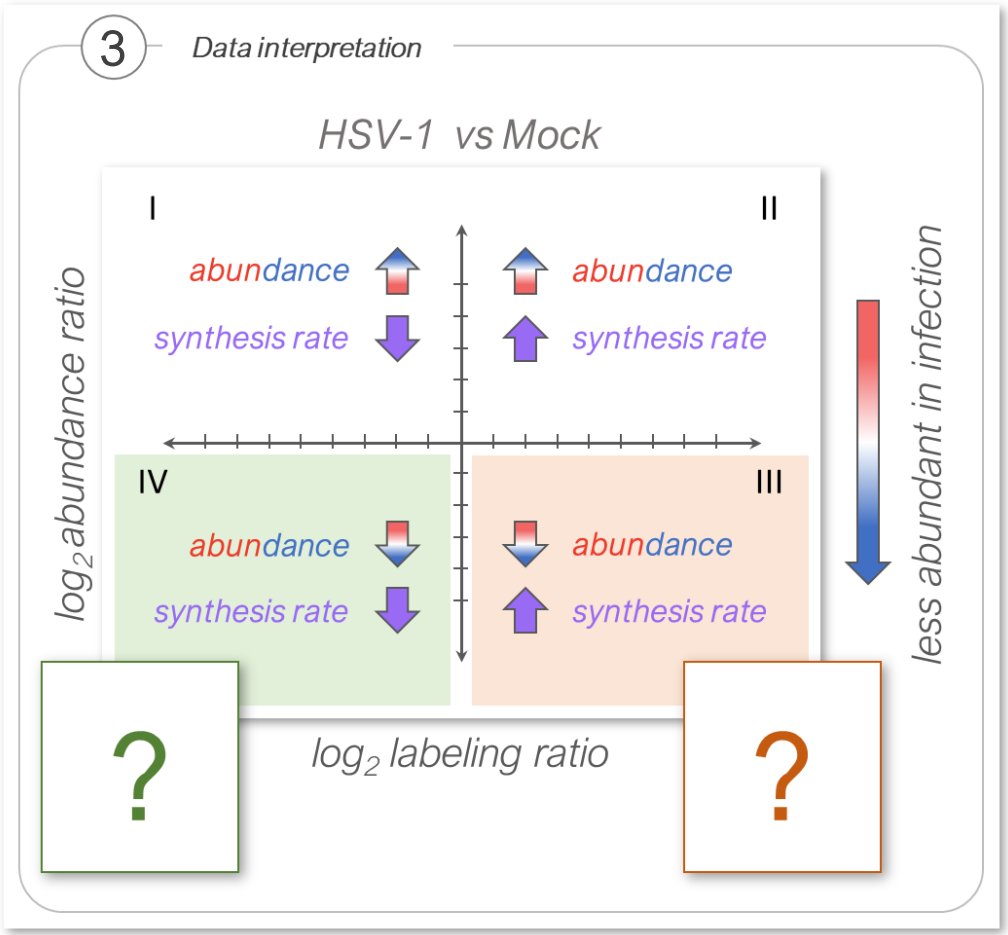


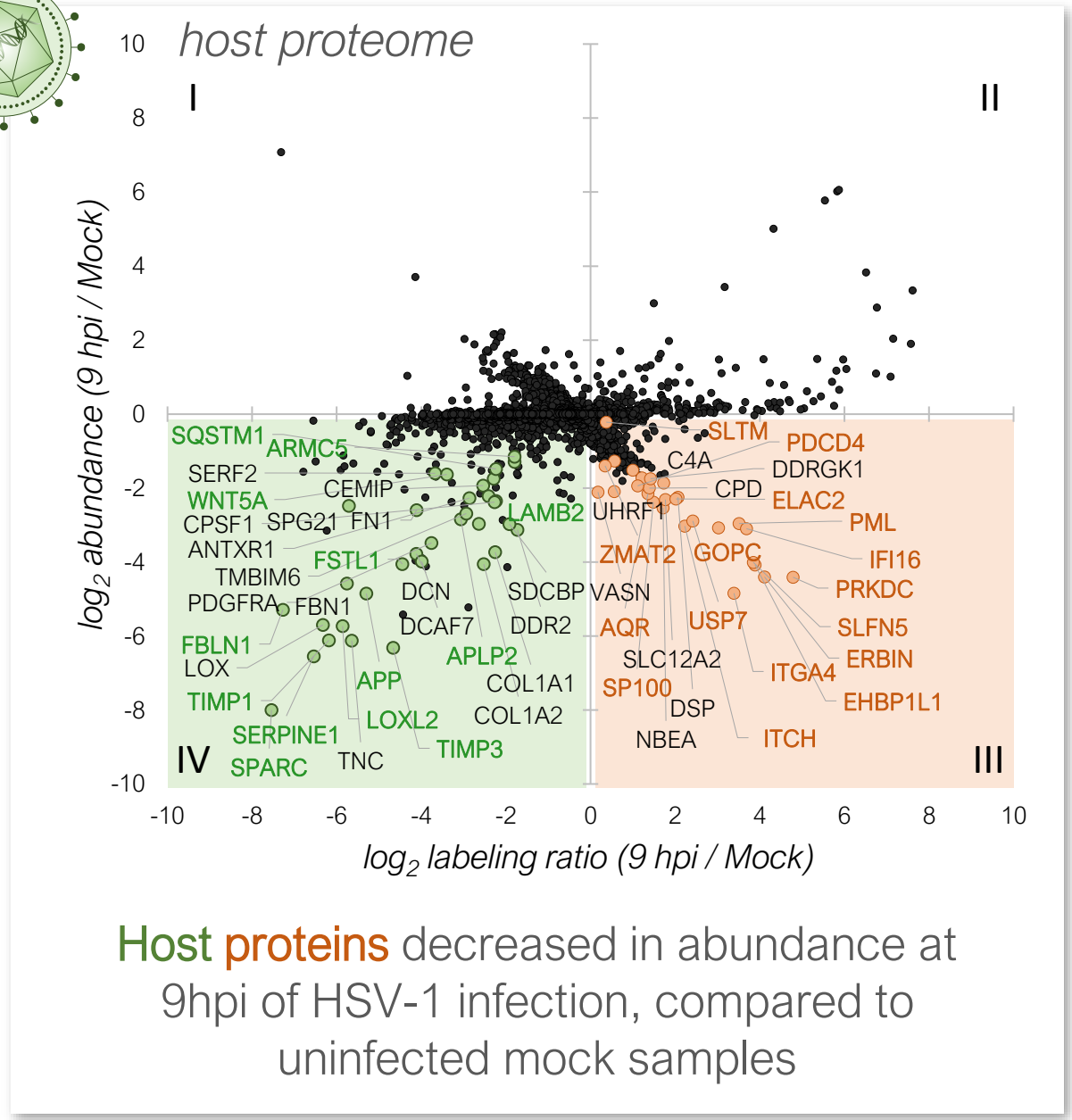
2 Data processing



Combine fold change and turnover

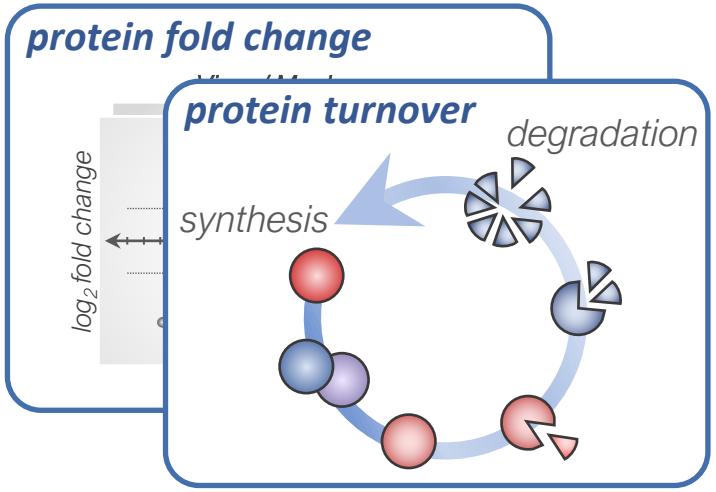


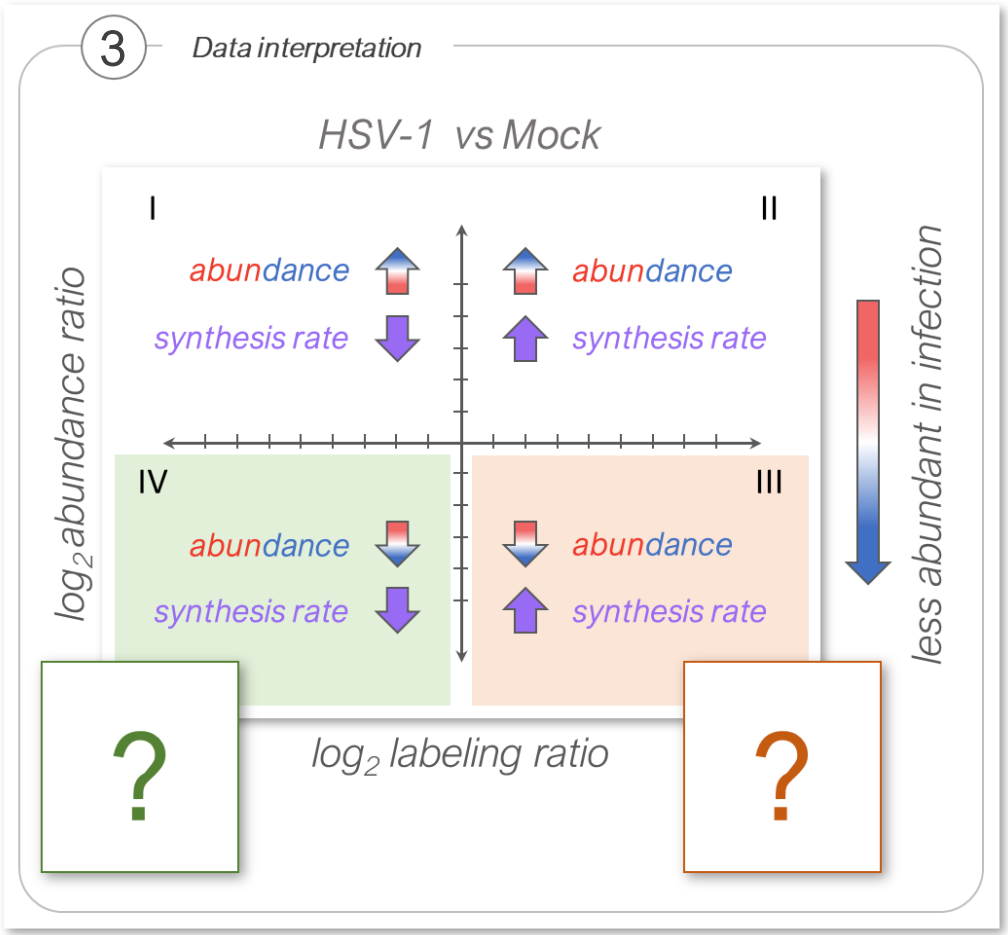


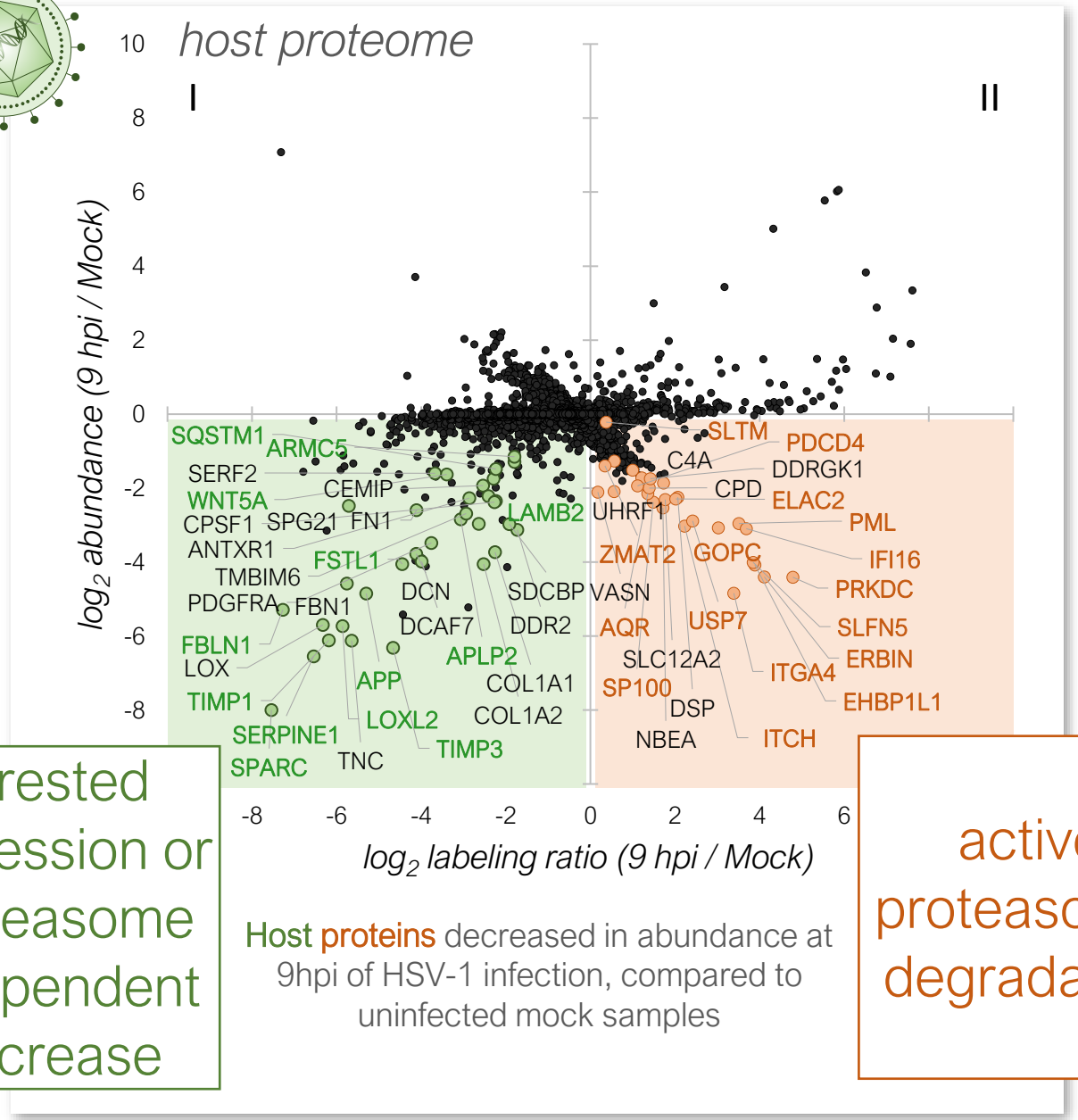


Host proteins decreased in abundance at 9hpi of HSV-1 infection, compared to uninfected mock samples

Combine fold change and turnover





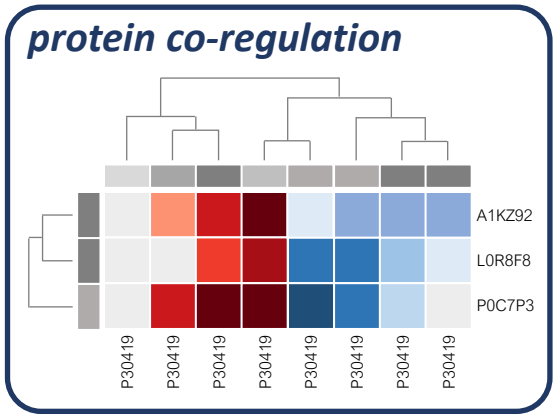


arrested expression or proteasome independent decrease

Host proteins decreased in abundance at 9hpi of HSV-1 infection, compared to uninfected mock samples

active proteasome-degradation

Other quantitative dimensions that assist the interpretation of protein interactions and functions

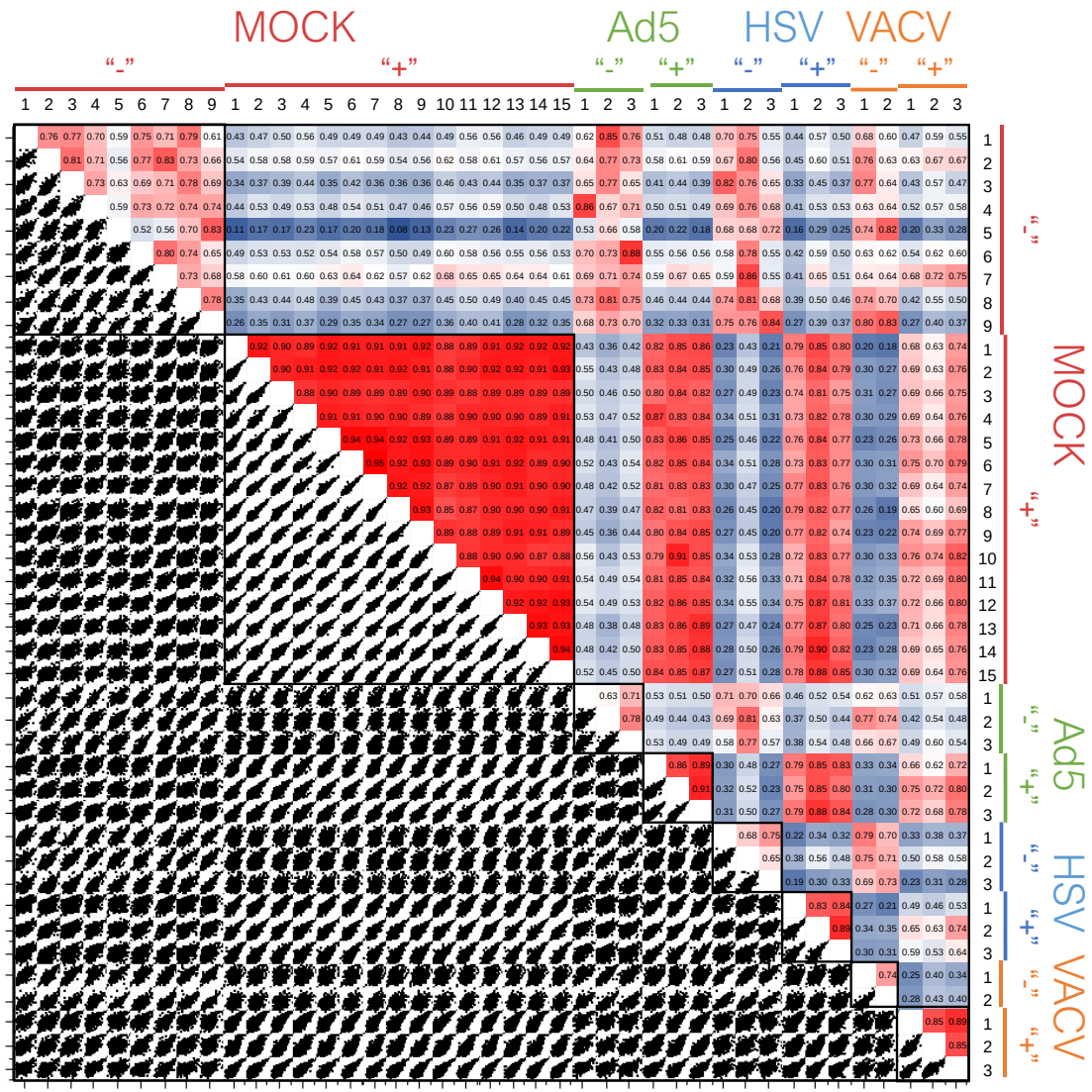


Co-regulation to identify genetic interactions

Ad5 – Adenovirus type-5
HSV – Herpes Simplex Virus
VACV – Vaccinia virus

Concept:

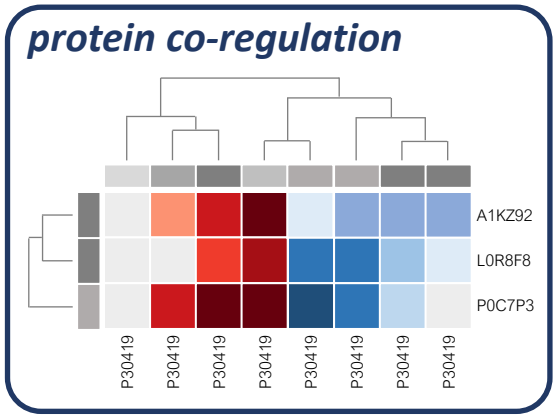
- *Proteins with the same regulation in a large set of samples have potentially shared functions.*



"-" biotin; "+" biotin

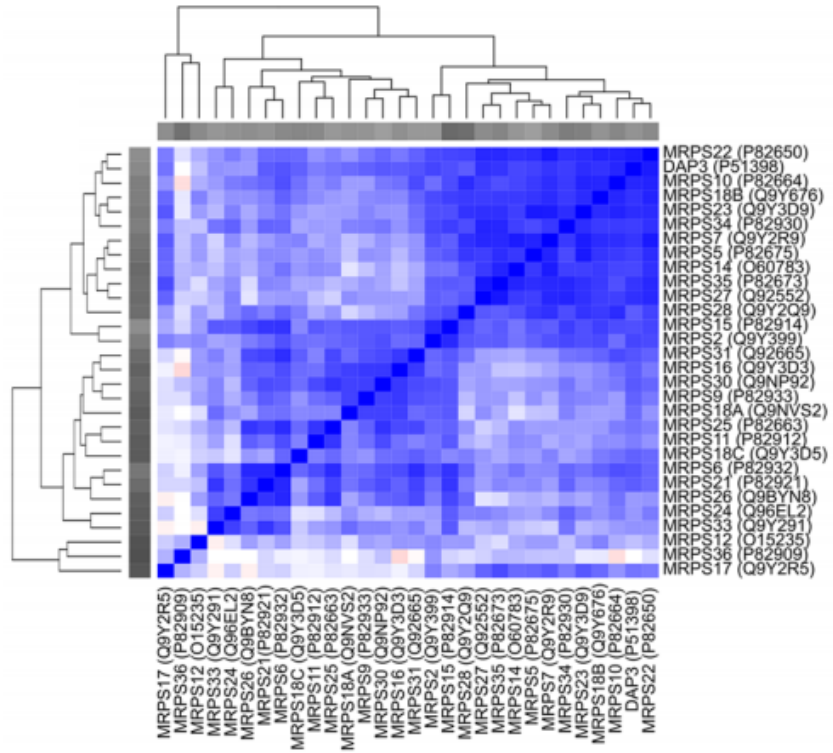
Reyes E.D. et al. (2017), Mol Cell Proteomics

Co-regulation to identify genetic interactions



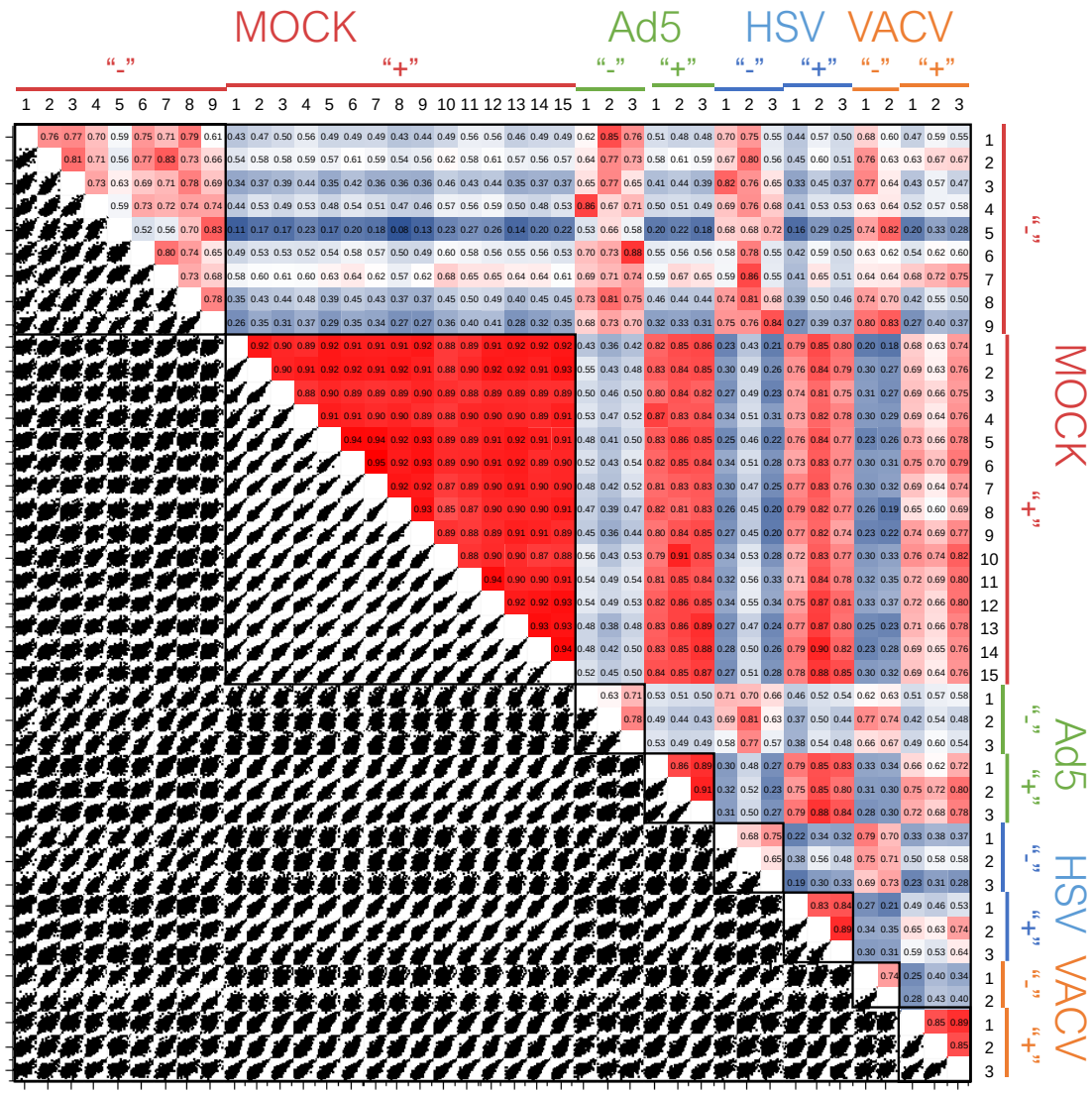
CoExpresso

Chalabi M.H., et al. (2019), BMC Bioinformatics.



- The analysis is based on protein abundance levels from ProteomicsDB. Correlations are shown as clustered heatmaps and as graphs to reveal potential co-regulated protein subgroups.

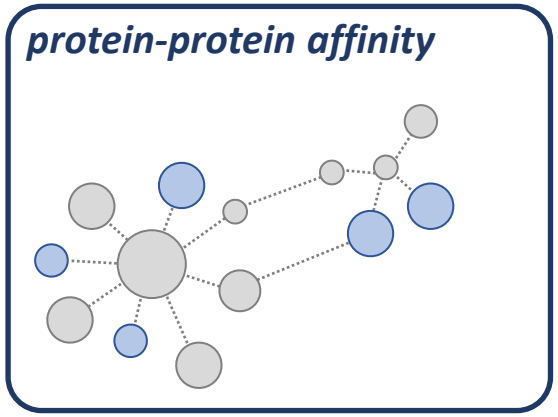
Ad5 –Adenovirus type-5 ; HSV – Herpes Simplex Virus; VACV – Vaccinia virus



"-" biotin; "+" biotin

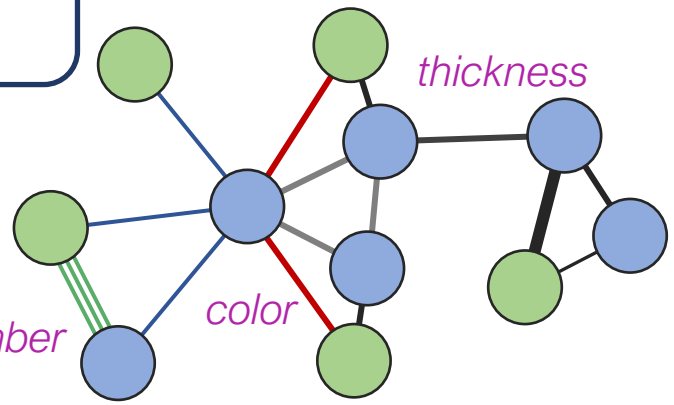
Reyes E.D. et al. (2017), Mol Cell Proteomics

Protein-protein interaction network

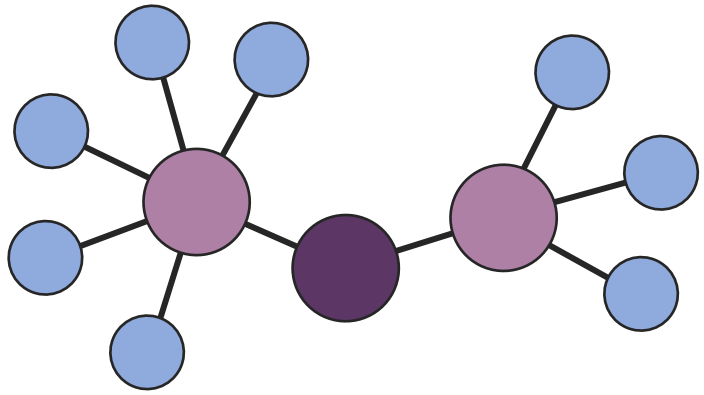


Common network types encountered in proteomics studies:

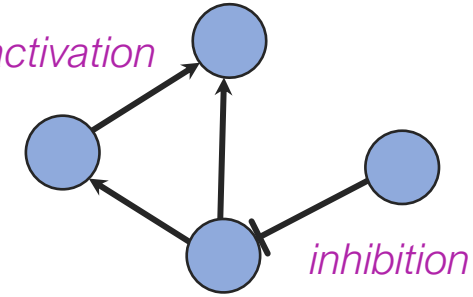
weighted



hierarchical



semantic pathway



In proteomics, most often, the nodes represent proteins, e.g.

- color of the node represents the fold change enrichment
- size - the p-value or abundance
- shape - the functional enrichments (e.g., Gene Ontology)

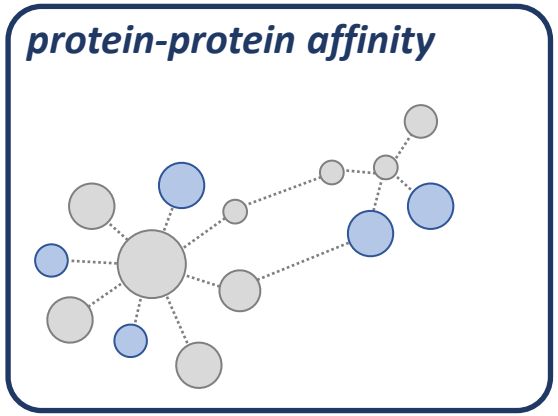
The edges usually represent one of three types of information:

- physical interaction or proximity (interactomics),
- phenotypic similarity (profiling),
- shared annotations (e.g., Gene Ontology)

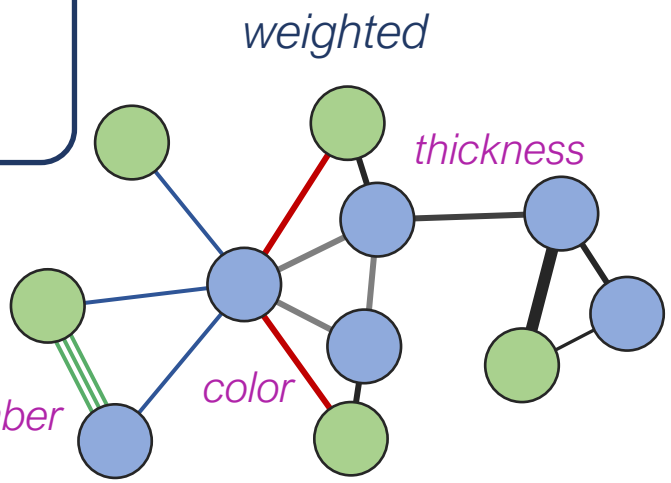
When creating a network, it is essential to understand the required network type, the key information that should be associated with nodes and edges, and how to represent it in the visualization when creating a network

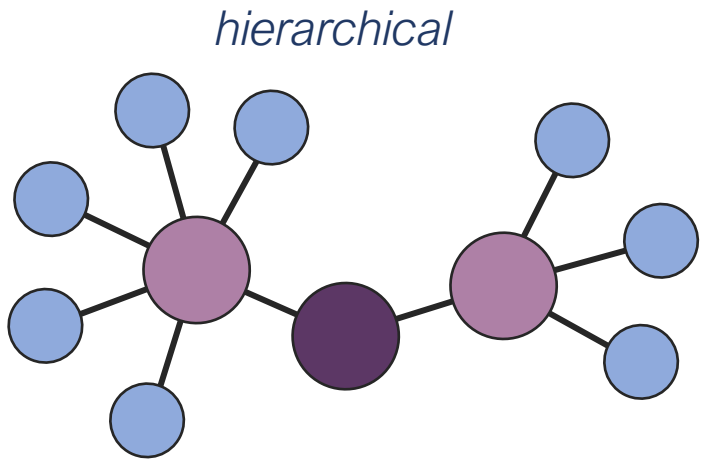
Schessner J.P., et al. (2022) Proteomics.
Koutrouli, M., et al. (2020), Front Bioeng Biotechnol.

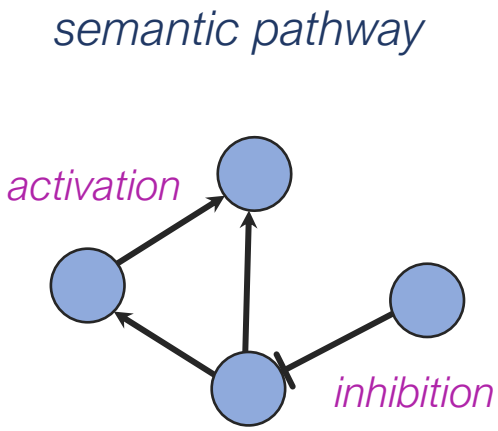
Protein-protein interaction network



Common network types encountered in proteomics studies:







Different tools can be used to create networks depending on the degree of complexity and customization required e.g.,:

Databases for protein-protein interaction networks:

- **STRING:** edges (connections) given by scores
- **BioGRID:** curated from publications (nearly complete for yeast)
- **KEGG:** more focused on pathways and metabolism
- **Reactome:** curated annotation of pathways on a diverse set of topics
- **NextProt:** an integrated collection of interactions between human proteins
- **IntAct:** curation from literature and submitted information

Databases for protein-PTMs interaction networks:

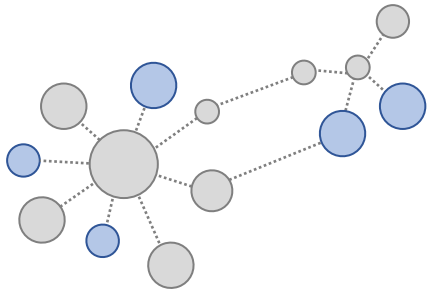
- **PhosphoSitePlus:** Eexperimentally reported ph-, ac- and ub-sites
- **UniProt:** experimental/putative PTMs
- **PhosphoPath:** visualization and analysis of quantitative proteome and phosphoproteome datasets
- **MetaCore:** data-mining and pathway analysis
- **Ingenuity IPA: a** tool to integrate and understand complex 'omics data

...endless number of specific databases

Schessner J.P., et al. (2022) Proteomics.

Protein-protein interaction network

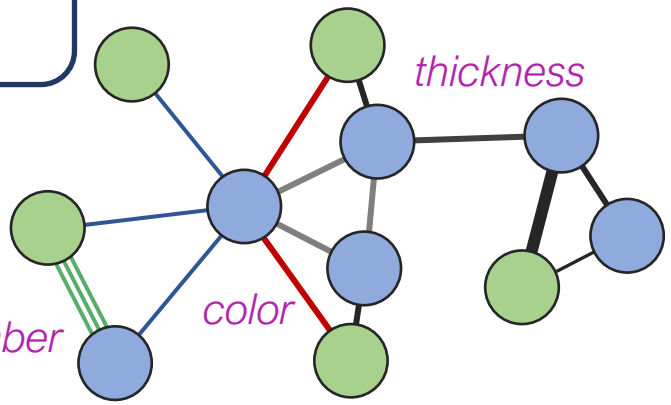
protein-protein affinity



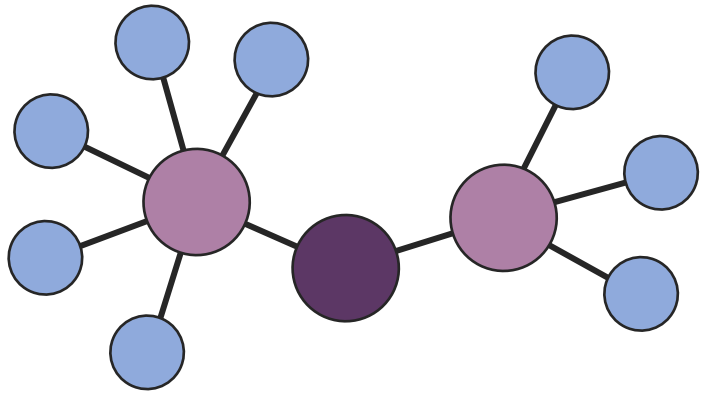
Cytoscape

Common network types encountered in proteomics studies:

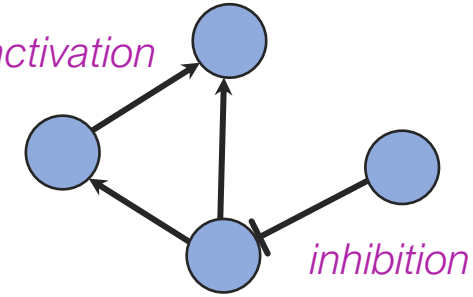
weighted



hierarchical

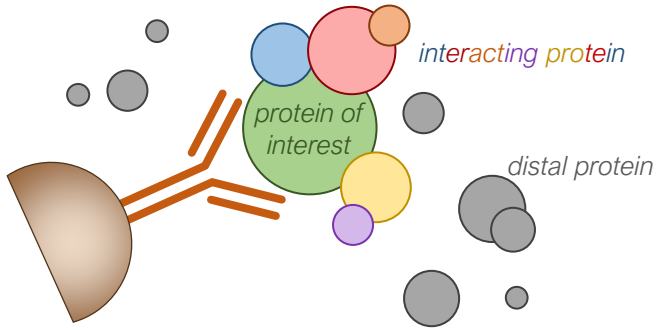


semantic pathway

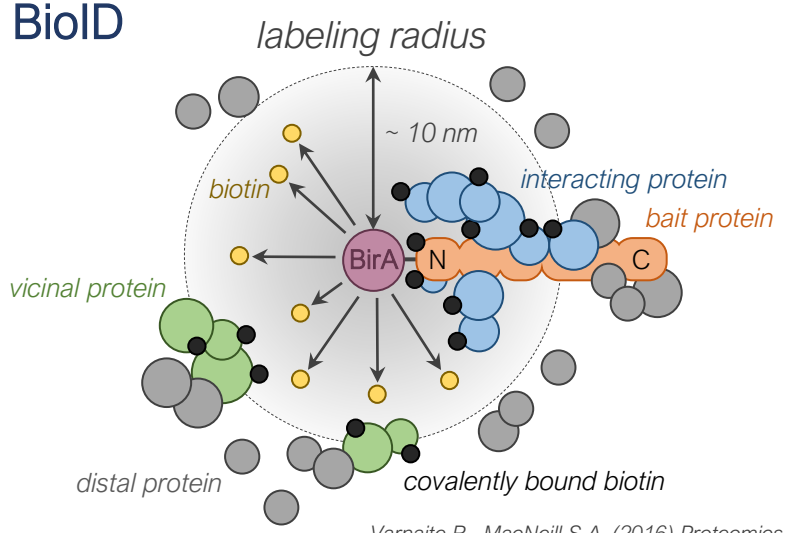


Protein-protein affinity is mostly performed by:

Immunoprecipitation

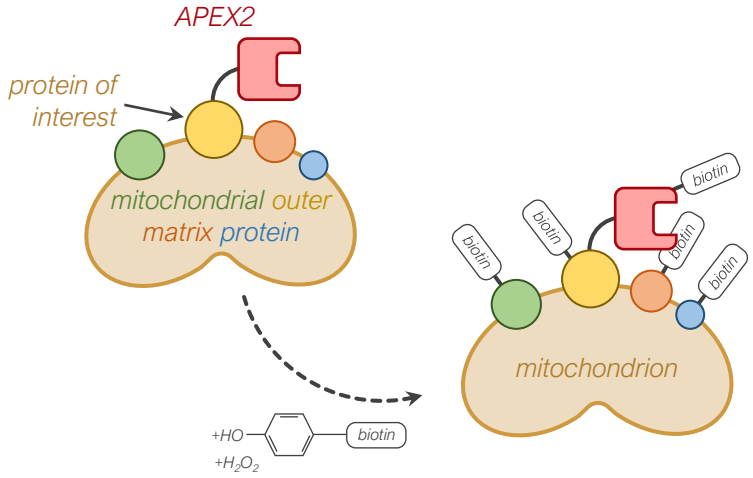


BioID



Varnaite R., MacNeill S.A. (2016) Proteomics.

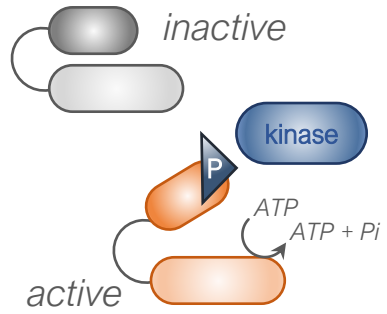
APEX IPs



Interaction AND PTMs

protein-protein affinity

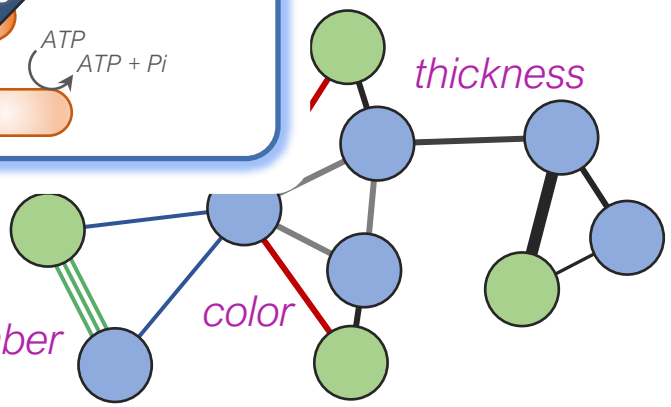
protein PTMs



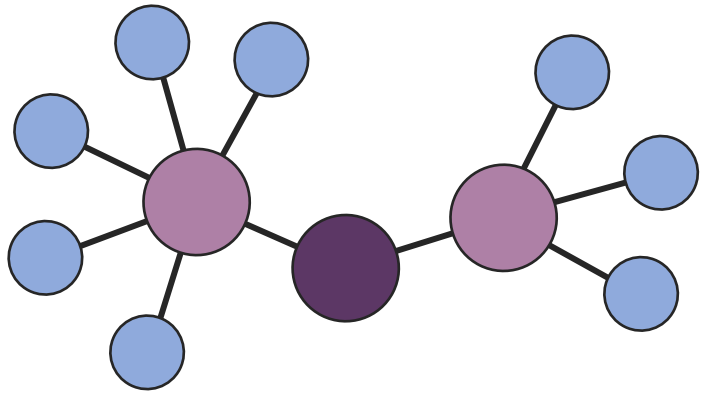
Cytoscape

Common network types encountered in proteomics studies:

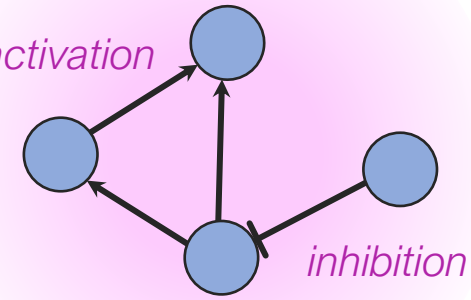
weighted



hierarchical

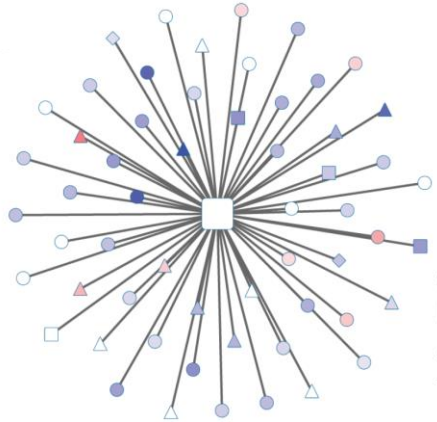


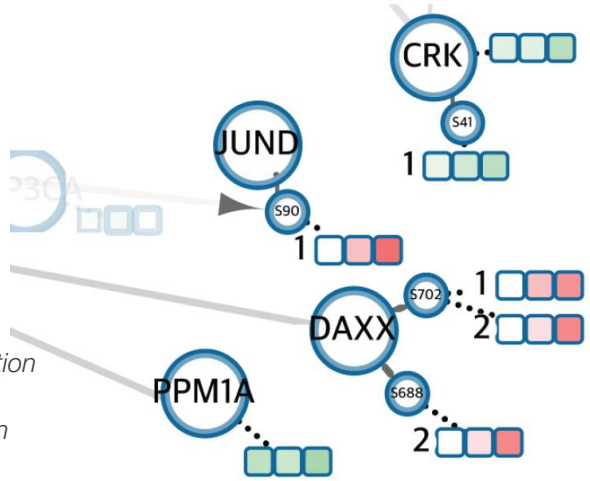
semantic pathway



Visualization and analysis of quantitative proteome and protein PTM datasets, e.g.:

PhosphoPath

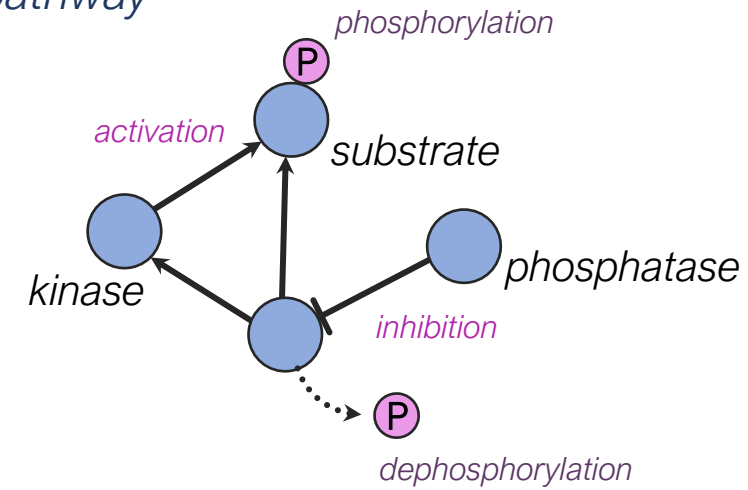


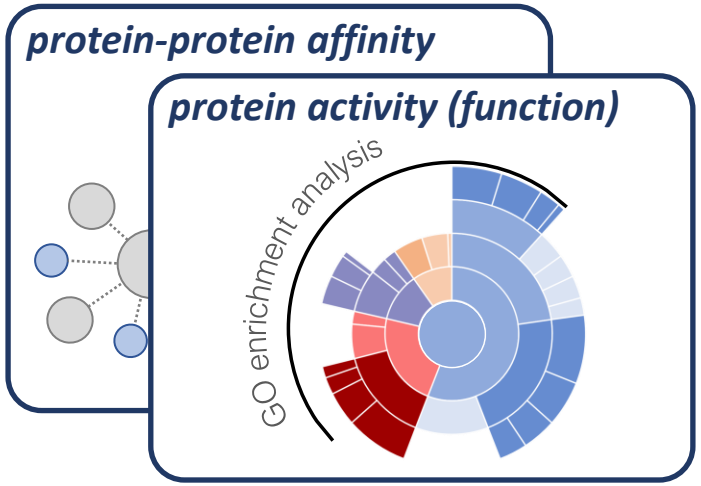


Phosphorylation
Acetylation
Ubiquitination
Methylation

https://apps.cytoscape.org/apps/phosphopath

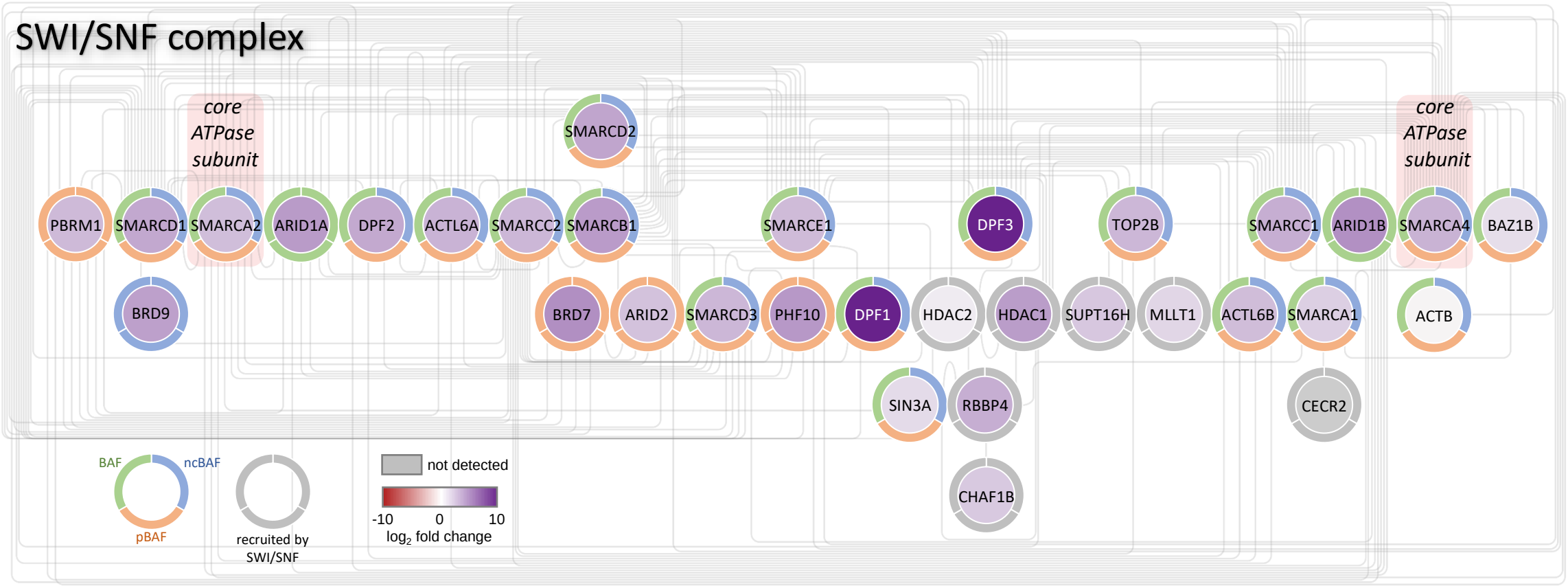
semantic pathway



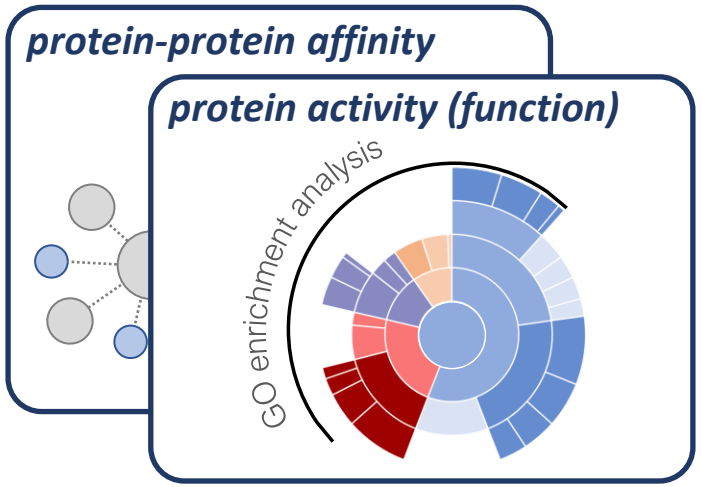


Interaction AND activity

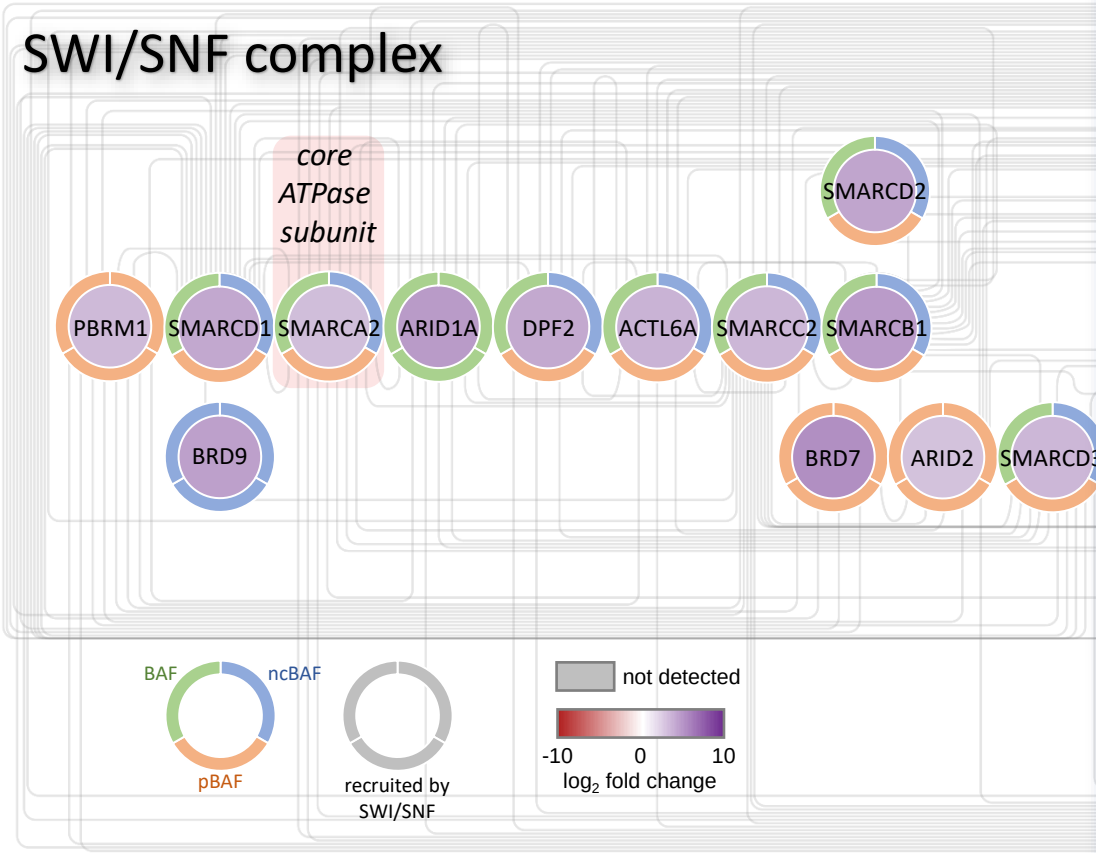
Layout apps available in Cytoscape

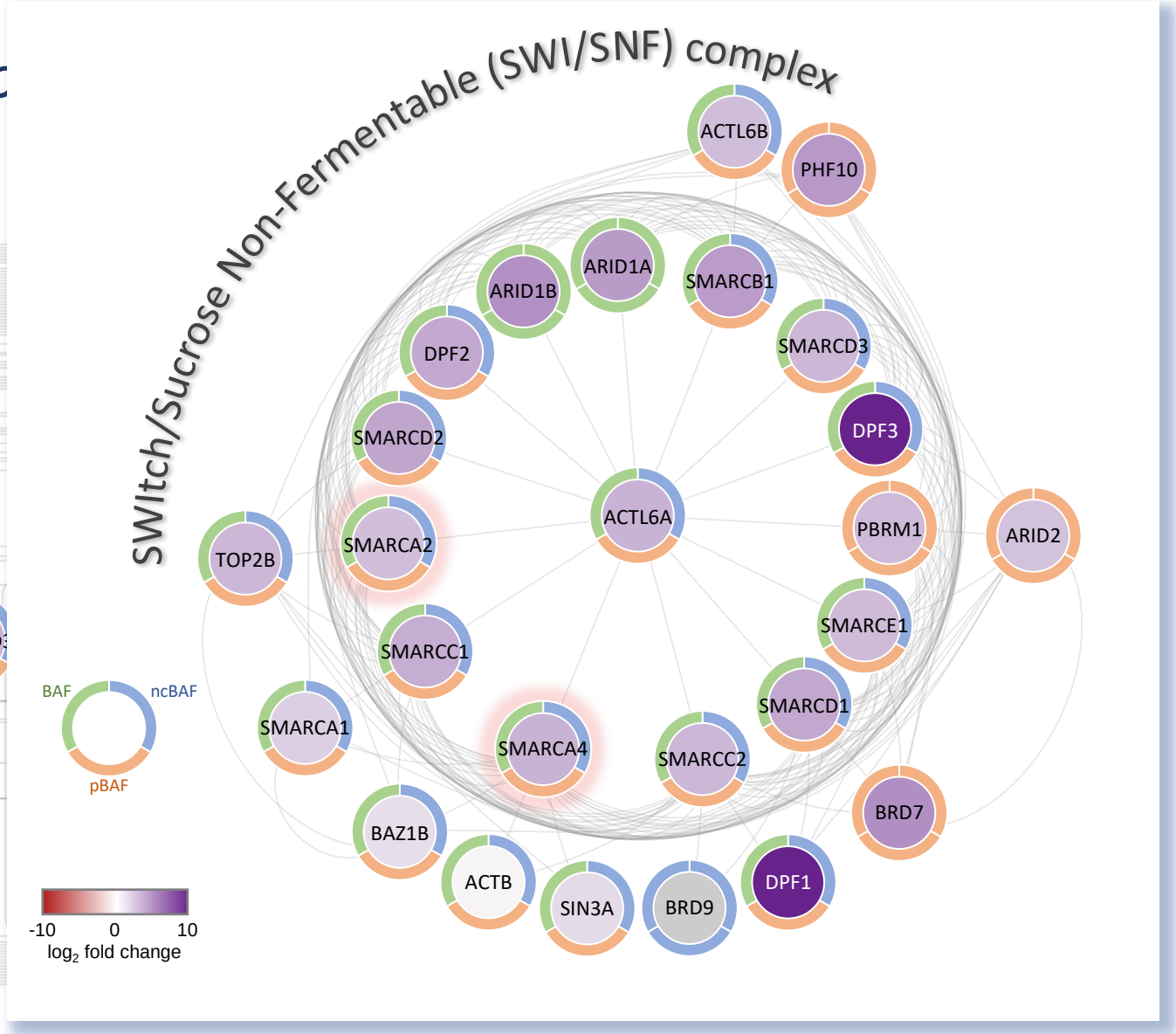


Interaction AND activity

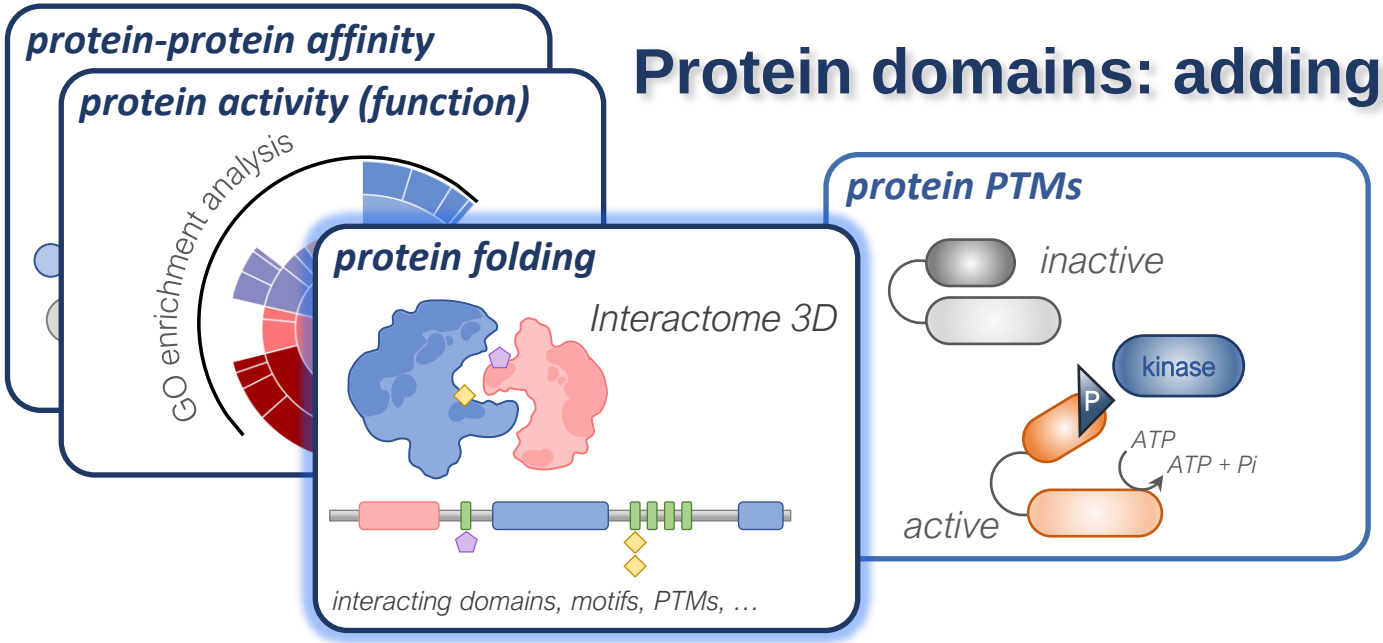


Layout ap





Protein domains: adding structural details to protein networks

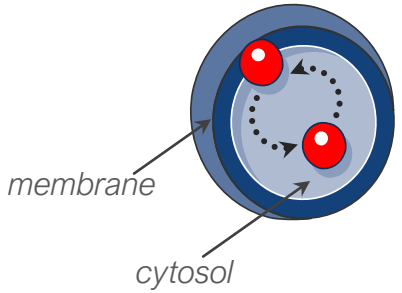


Motifs/domains

- Prosite: Biologically significant sites, patterns and profiles
  Pfam: Collection of protein domain families
  Blocks: Database of conserved protein "blocks" (short sequences)
  PRINTS: Protein fingerprints (groups of conserved motifs)
- InterPro: protein sequence analysis & classification - searches several databases
- ScanProsite: consists of documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles to identify them
- SMART: (a Simple Modular Architecture Research Tool) allows the identification and annotation of genetically mobile domains and the analysis of domain architectures.

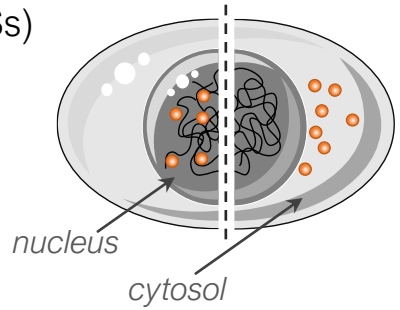
Membrane binding domains e.g.,

- C1,
- C2,
- PH,
- FYVE,
- PX,
- ENTH,
- and BAR domains



The transport of molecules between the nucleus and the cytosol e.g.,

- nuclear localization signals (NLSs)
- ❑ export peptide signal
- ❑ import peptide signal

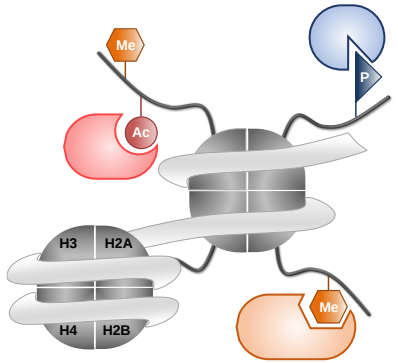


Protein secretion e.g.,

- specific peptide tags

Protein PTMs binding domains e.g.,

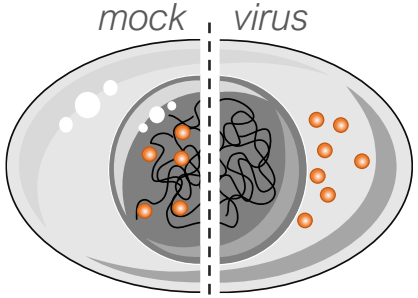
- 14-3-3,
- WW domains,
- LRR domains of F-box proteins,
- FHA domains
- Bromodomains
- methyl-CpG-binding domain (MBD)



Protein PTMs AND translocation

protein PTMs

protein translocation



Rat brain homogenate

centrifugation

P1: Pellet

S1: Supernatant

centrifugation

Subcellular fractionation

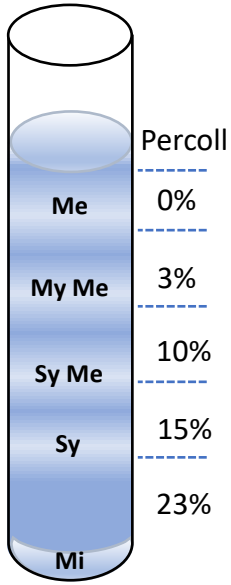
F1: Membranes

F2:Myelin and membranes

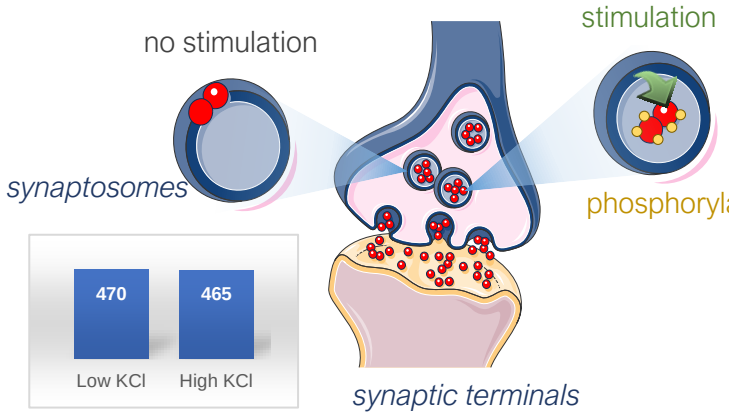
F3:Synaptosomes and membranes

F4: Synaptosomes

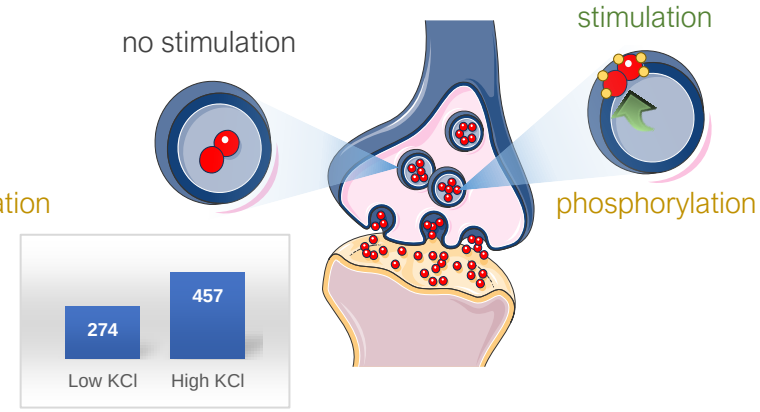
F5: Mitochondria



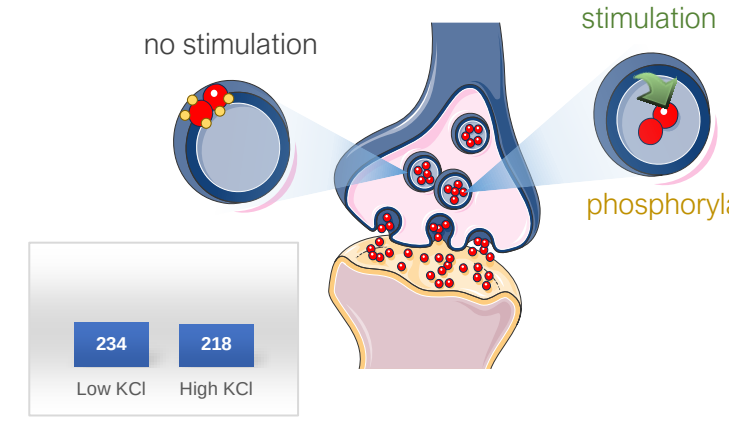
Upon phosphorylation translocating to the cytosol



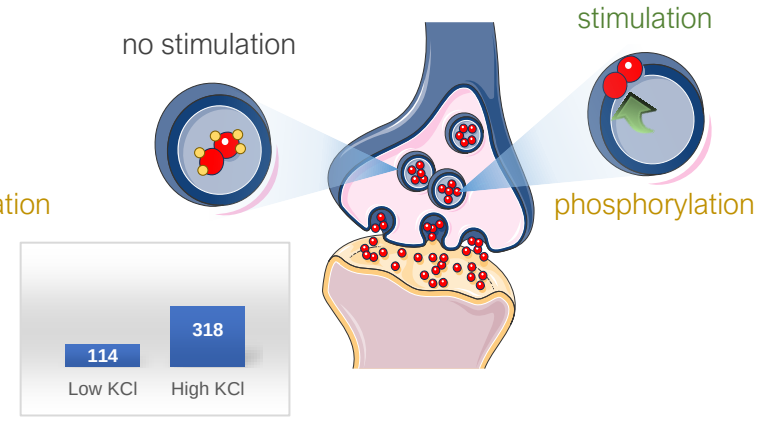
Upon phosphorylation translocating to the membrane



Upon dephosphorylation translocating to the cytosol



Upon dephosphorylation translocating to the membrane



Conclusions

- We hope we provided an original overview of ideas for experimental design
- Data produced by mass spectrometry provide perspectives beyond fold changes
- Exploiting labeling, time series, modifications and large resources of available data can be integrated in single analyses to resolve multiple properties of the regulated proteome
- As well, co-regulation, rank of abundance, and protein 3D structure are poorly exploited aspects for data interpretation. We hope this works as a light refreshment prior to your evening fun ☺

MetaboLights & Repository Level Workflows

Leveraging computational resource and new annotation logic to draw fresh insight

Callum Martin

Software Developer

ebi.ac.uk/metabolights

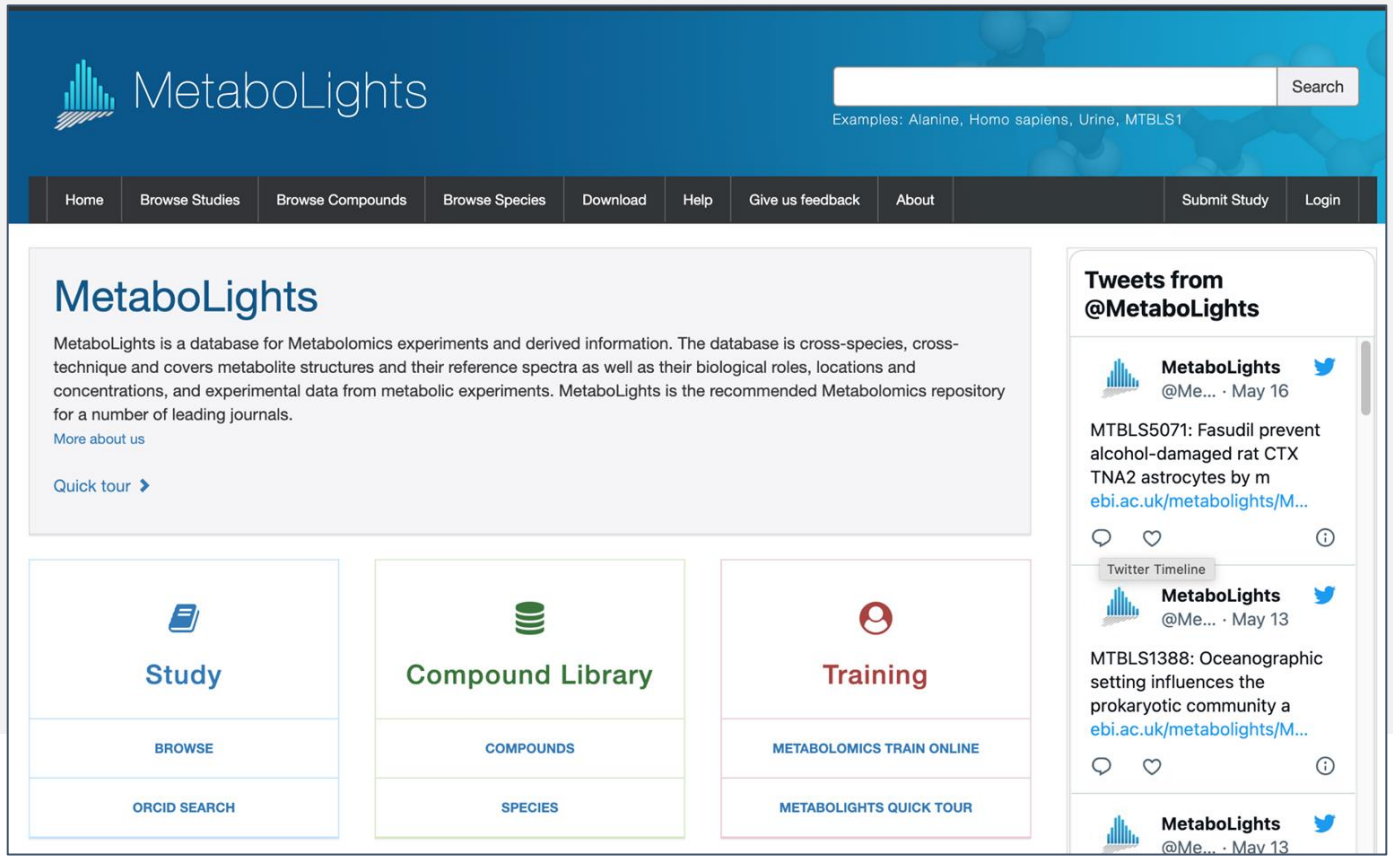
EMBL-EBI

What is MetaboLights?

MetaboLights

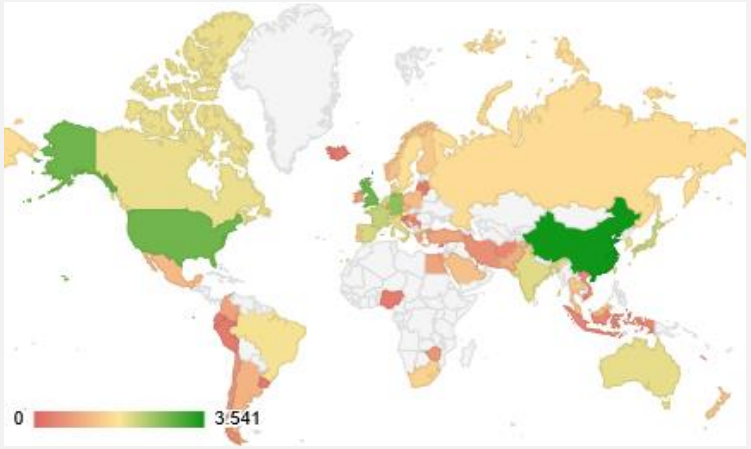
Open Source Study Repository & Metabolite Knowledgebase

The database is **cross-species**, **cross-technique** and covers **metabolite structures and their reference spectra** as well as their biological roles, locations and concentrations, and **experimental data** from metabolic experiments



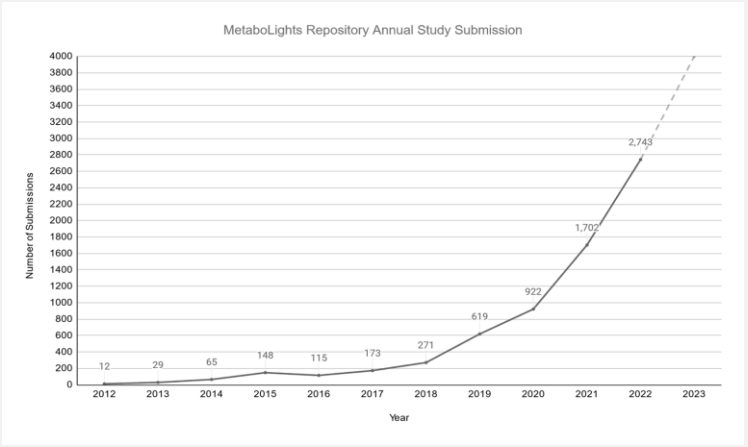
https://www.ebi.ac.uk/metabolights

Submissions

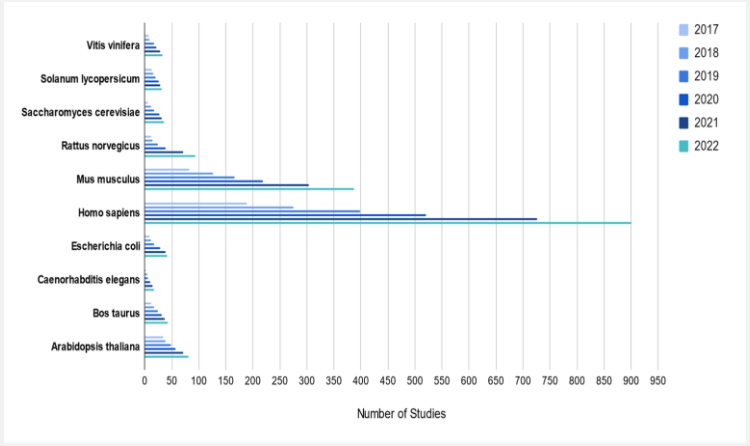


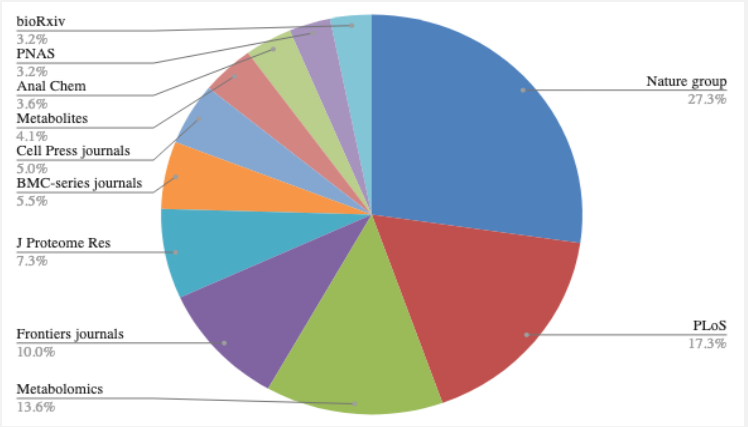
Geographical distribution of submitted studies

(Top 10: China: 1490, USA: 509, UK: 450, Germany: 309, France: 124, Japan: 102, Spain: 87, Italy: 80, Australia: 71, India: 66)



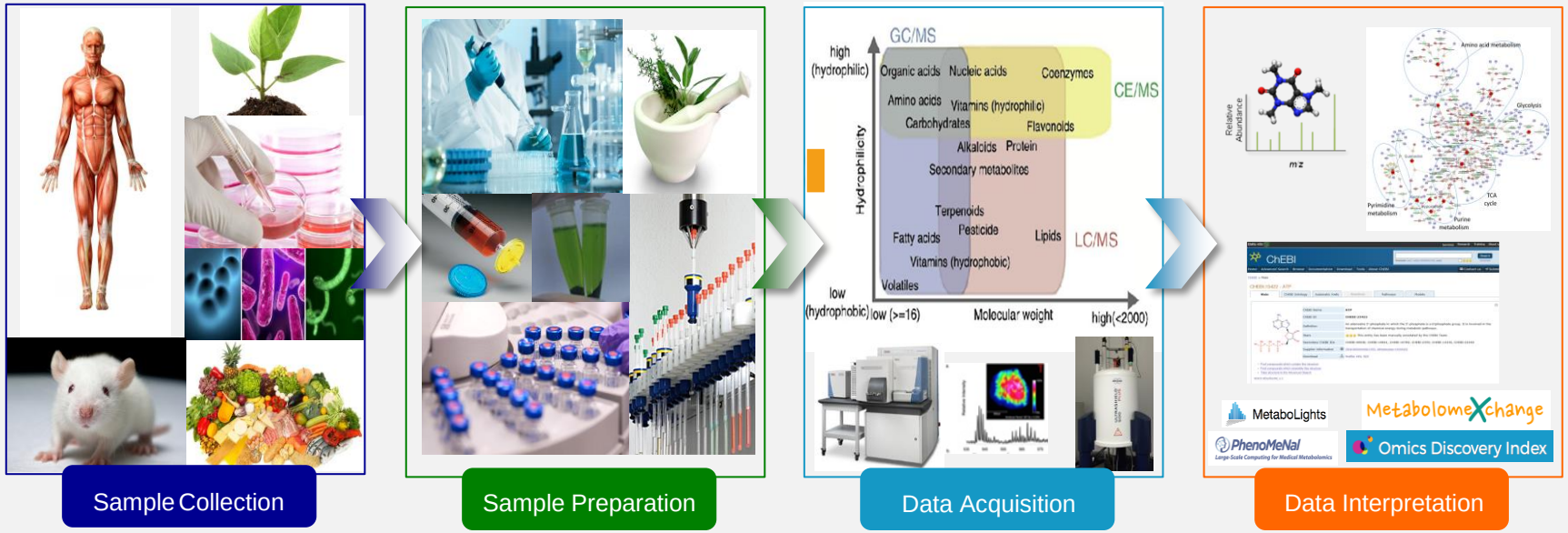
MetaboLights study submission rates per year





What exactly does our data look like?

Broadly we have two categories of interest - metadata, the information that describes the experiment, and raw and derived data / files, the outputs from the experiment



Status Public | Release Date 2019-09-19 | Edit | ISA

MTBLS749: Alterations in the tyrosine and phenylalanine pathways revealed by biochemical profiling in cerebrospinal fluid of Huntington's disease subjects

Kim Kultima, Stephanie Herman

Huntington's disease (HD) is a severe neurological disease leading to psychiatric symptoms, motor impairment and cognitive decline. The disease is caused by a CAG expansion in the huntingtin (HTT) gene, but how this translates into the clinical phenotype of HD remains elusive. Using liquid chromatography mass spectrometry, we analyzed the metabolome of cerebrospinal fluid (CSF) from premanifest and manifest HD subjects as well as control subjects. Inter-group differences revealed that the tyrosine metabolism, including tyrosine, thyroxine, L-DOPA and dopamine, was significantly altered in manifest compared with premanifest HD. These metabolites demonstrated moderate to strong associations to measures of disease severity and symptoms. Thyroxine and dopamine also correlated with the five year risk of onset in premanifest HD subjects. The phenylalanine and the purine metabolisms were also significantly altered, but associated less to disease severity. Decreased levels of lumichrome were commonly found in mutated HTT carriers and the levels correlated with the five year risk of disease onset in premanifest carriers. These biochemical findings demonstrates that the CSF metabolome can be used to characterize molecular pathogenesis occurring in HD, which may be essential for future development of novel HD therapies.

PUBLICATIONS

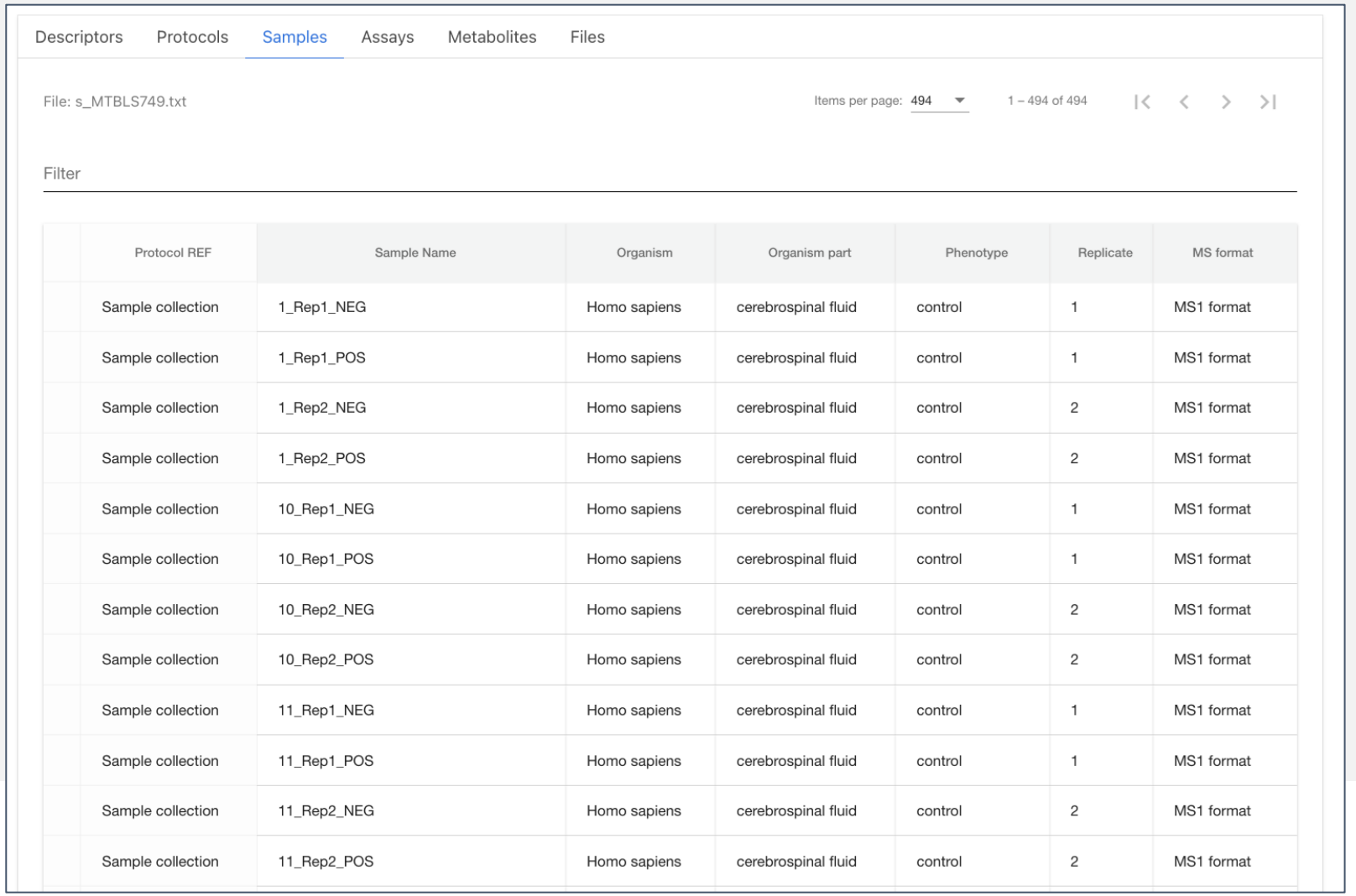
Alterations in the tyrosine and phenylalanine pathways revealed by biochemical profiling i...

Herman Stephanie, Niemelä Valter, Emami Khoonsari Payam, Sun...

Descriptors | Protocols | Samples | Assays | Metabolites | Files | Validations

Sample Information

Information pertaining to each individual sample processed in a study



Descriptors | Protocols | Samples | Assays | Metabolites | Files

File: s_MTBLS749.txt

Items per page: 494 | 1 – 494 of 494

Filter

Protocol REF	Sample Name	Organism	Organism part	Phenotype	Replicate	MS format
Sample collection	1_Rep1_NEG	Homo sapiens	cerebrospinal fluid	control	1	MS1 format
Sample collection	1_Rep1_POS	Homo sapiens	cerebrospinal fluid	control	1	MS1 format
Sample collection	1_Rep2_NEG	Homo sapiens	cerebrospinal fluid	control	2	MS1 format
Sample collection	1_Rep2_POS	Homo sapiens	cerebrospinal fluid	control	2	MS1 format
Sample collection	10_Rep1_NEG	Homo sapiens	cerebrospinal fluid	control	1	MS1 format
Sample collection	10_Rep1_POS	Homo sapiens	cerebrospinal fluid	control	1	MS1 format
Sample collection	10_Rep2_NEG	Homo sapiens	cerebrospinal fluid	control	2	MS1 format
Sample collection	10_Rep2_POS	Homo sapiens	cerebrospinal fluid	control	2	MS1 format
Sample collection	11_Rep1_NEG	Homo sapiens	cerebrospinal fluid	control	1	MS1 format
Sample collection	11_Rep1_POS	Homo sapiens	cerebrospinal fluid	control	1	MS1 format
Sample collection	11_Rep2_NEG	Homo sapiens	cerebrospinal fluid	control	2	MS1 format
Sample collection	11_Rep2_POS	Homo sapiens	cerebrospinal fluid	control	2	MS1 format

Assay Information

Descriptors | Protocols | Samples | Assays | Metabolites | Files

Assay Sheet 1 | Assay Sheet 2 | Assay Sheet 3 | Assay Sheet 4

File: a_MTBLS1987_LC-MS_positive_hilic_metabolite_profiling.txt

Items per page: 69 | 1 – 69 of 69

Filter

Protocol REF	Parameter Value - Chromatography Instrument	Parameter Value - Autosampler model	Parameter Value - Column model	Parameter Value - Column type	Parameter Value - Guard column	Labeled Extract Name	Label	Protocol REF	Parameter Value - Scan polarity	Parameter Value - Scan m/z range	Parameter Value - Instrument
Chromatography	Waters ACQUITY UPLC H-Class System		ACQUITY UPLC BEH Amide (1.7 µm, 2.1 mm x 100 mm; Waters)	HILIC				Mass spectrometry	Positive	50-1000	Waters Xevo G2-S QTof
Chromatography	Waters ACQUITY UPLC H-Class System		ACQUITY UPLC BEH Amide (1.7 µm, 2.1 mm x 100 mm; Waters)	HILIC				Mass spectrometry	Positive	50-1000	Waters Xevo G2-S QTof
Chromatography	Waters ACQUITY UPLC H-Class System		ACQUITY UPLC BEH Amide (1.7 µm, 2.1 mm x 100 mm; Waters)	HILIC				Mass spectrometry	Positive	50-1000	Waters Xevo G2-S QTof

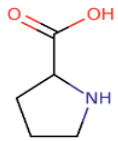
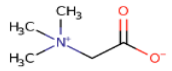
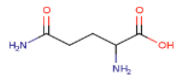
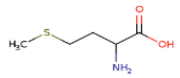
Metabolite Annotation Information

Descriptors | Protocols | Samples | Assays | Metabolites | Files

MAF Sheet 1 | MAF Sheet 2 | MAF Sheet 3 | MAF Sheet 4

File: m_MTBLS1987_LC-MS_positive_hilic_metabolite_profiling_v2_maf.tsv Items per page: 10 1 – 10 of 13

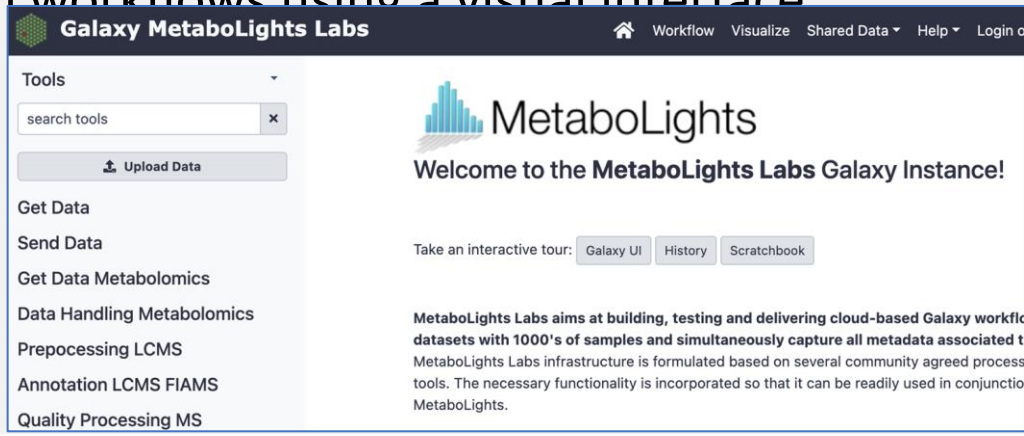
Filter

Structure	Database identifier	Chemical formula	SMILES	InChI	Metabolite identification	Mass to charge	Fra
	CHEBI:26271		OC(=O)C1CCCN1	InChI=1S/C5H9NO2/c7-5(8)4-2-1-3-6-4/h4,6H,1-3H2,(H,7,8)	Proline	116.07029	
	CHEBI:17750		C[N+](C)(C)CC([O-])=O	InChI=1S/C5H11NO2/c1-6(2,3)4-5(7)8/h4H2,1-3H3	Betaine	118.08747	
	CHEBI:28300		NC(CCC(N)=O)C(O)=O	InChI=1S/C5H10N2O3/c6-3(5(9)10)1-2-4(7)8/h3H,1-2,6H2,(H2,7,8)(H,9,10)	Glutamine	147.07607	
	CHEBI:16811		CSCCC(N)C(O)=O	InChI=1S/C5H11NO2S/c1-9-3-2-4(6)5(7)8/h4H,2-3,6H2,1H3,(H,7,8)	Methionine	150.05818	

MetaboLights Data & Galaxy Workflows

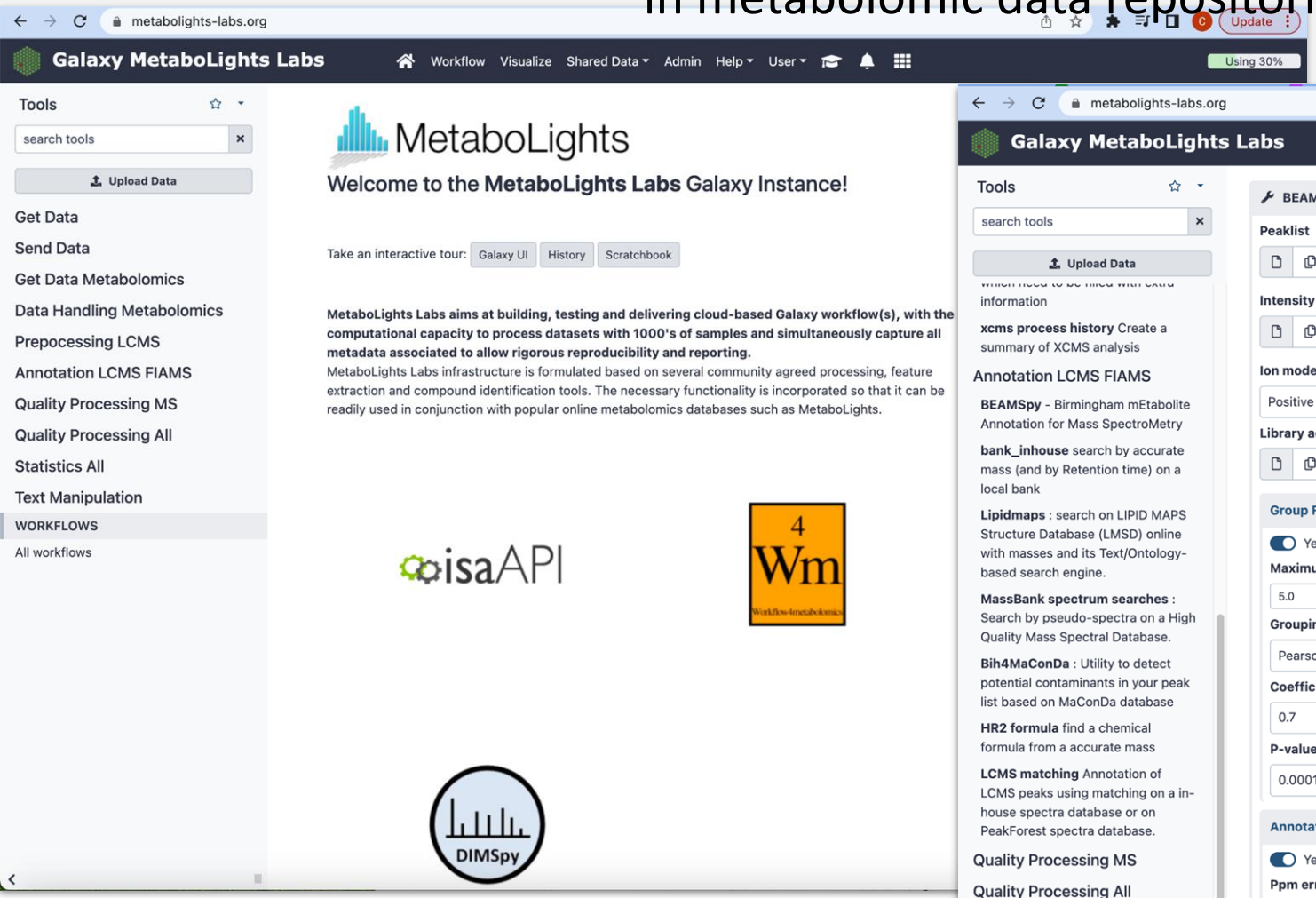
Galaxy is an open source web based computational platform that facilitates data-intensive biomedical research and analysis

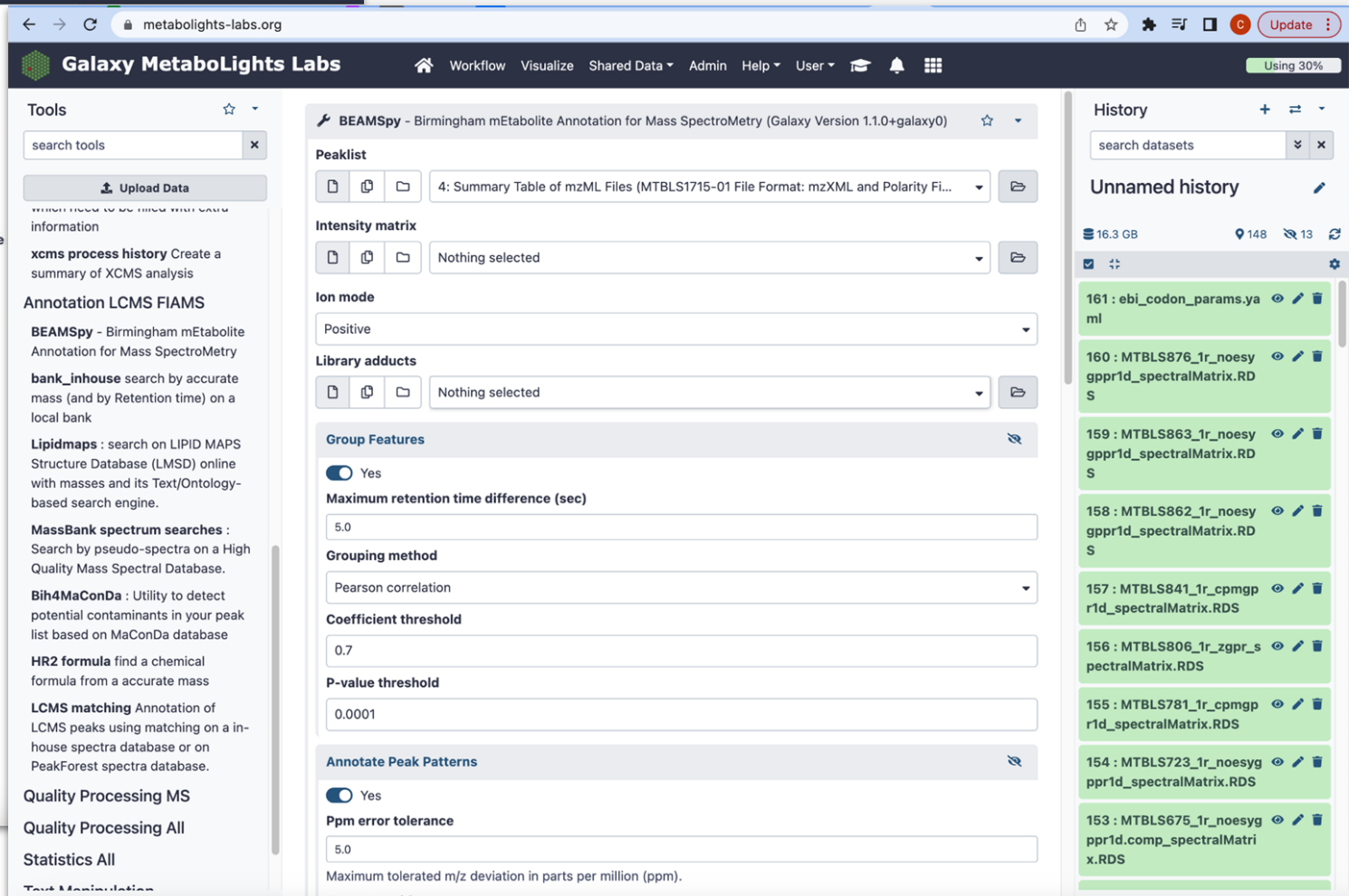
- It provides a user-friendly interface for data-intensive biomedical research and analysis
- It supports reproducible research by capturing and recording the entire analysis process
- Galaxy integrates a comprehensive collection of bioinformatics and data analysis tools
- Researchers can create and execute complex computational workflows using a visual interface
- We have our own instance at metabolights-labs.org
- (but you can deploy it anywhere including locally)



MetaboLights Workflows

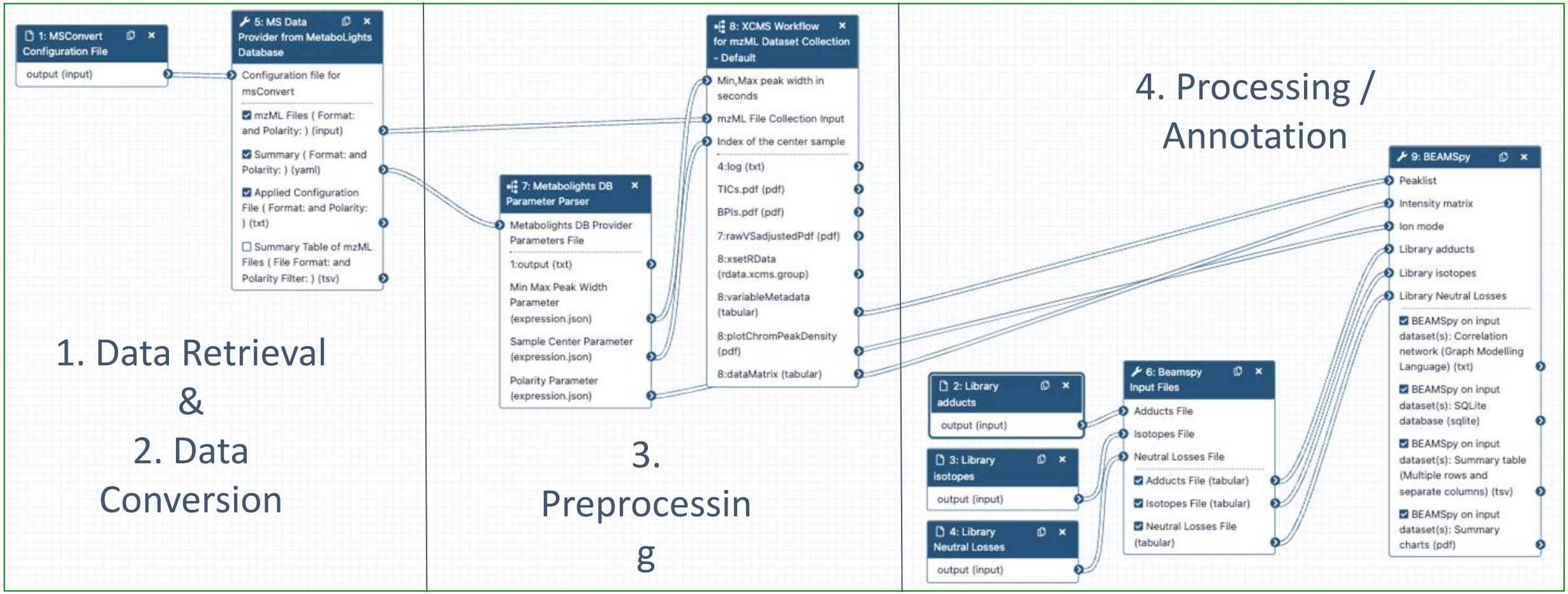
"Standardised metabolite annotation workflows for enhancing biological interpretation in metabolomic data repositories"





https://metabolights-labs.org

MetaboLights Labs - Workflows - MDP + XCMS + BEAMSpy

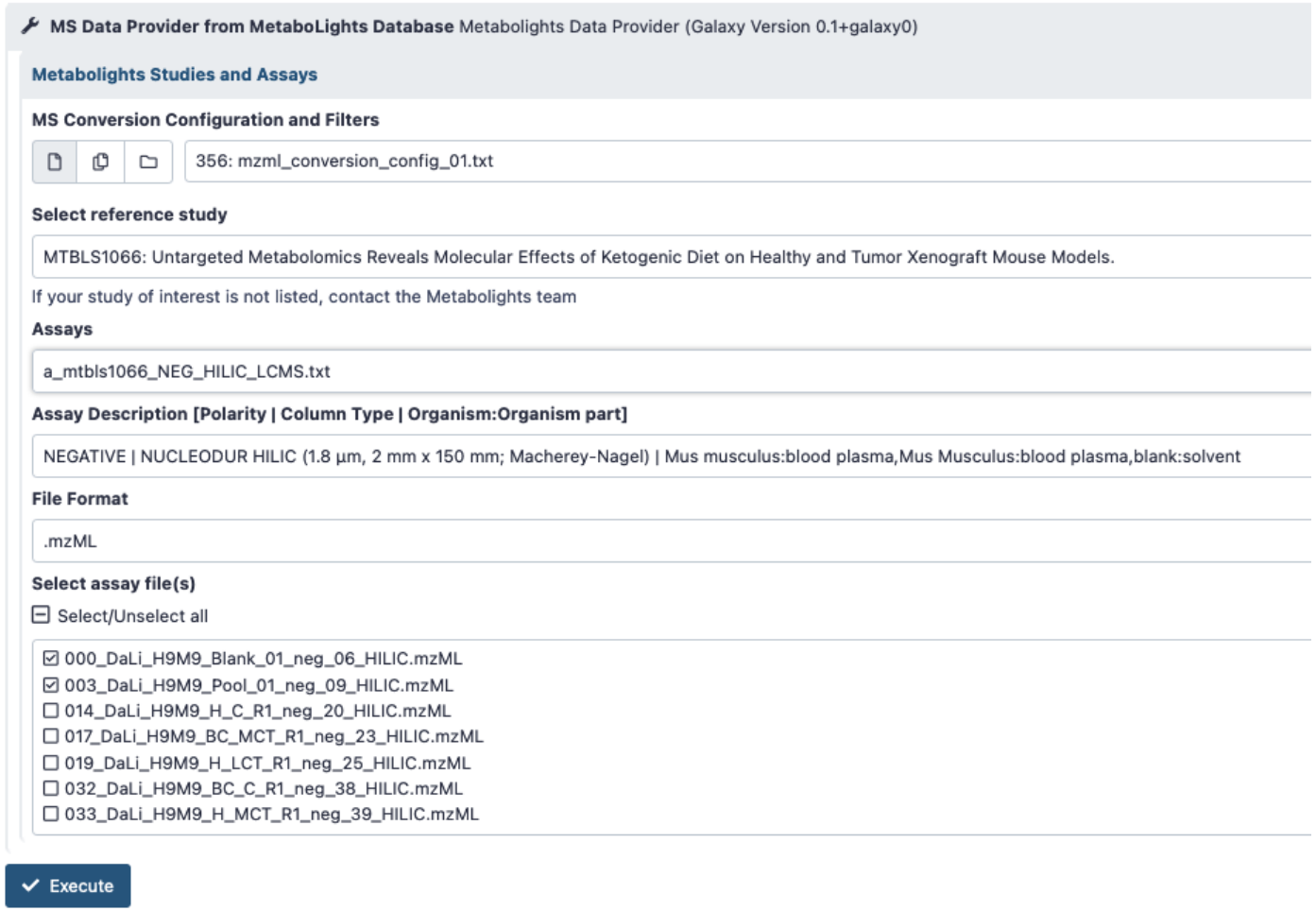


Rick Dunn (PI),
Tim Ebbels (PI),
Claire O'Donovan (PI)

UNIVERSITY OF LIVERPOOL
Imperial College London
UNIVERSITY OF BIRMINGHAM
EMBL-EBI

MetaboLights Data Provider (MDP) Tool

- Researchers can select files from MetaboLights and apply custom msConvert filters
  - Studies: 596, Assays: 1271, Assay Files: 174565
- Creates an open-source MS standard / file for preprocessing



proteowizard

MetaboLights Data Provider (MDP) Tool

Organism & Organism Parts & File Formats

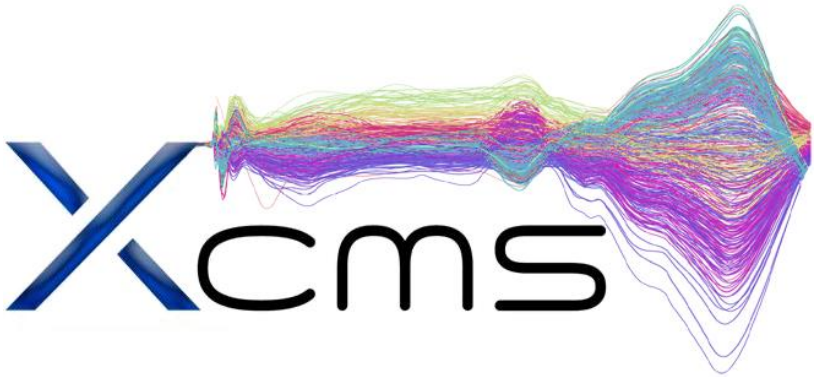
Organisms	Count of files
Homo sapiens	57987
Lolium perenne	23914
Mus musculus	7355
Saccharomyces cerevisiae	5984
reference compound	4623
blank	2717
sea water	2654
Vitis vinifera	2193
Ovis aries	1948
Rattus norvegicus	1943
groundwater	1859
Camellia sinensis	1788
Solanum lycopersicum x Solanum pennellii	1694
Gallus gallus	1562
Arabidopsis thaliana	1465
Bos taurus	1427
Chenopodium quinoa	1407
Escherichia coli	1326
Drosophila melanogaster	1047
...	

Organism Parts	Count of files
leaf	28844
blood serum	16528
blood plasma	15909
whole organism	10813
urine	6445
feces	3754
exometabolome	3471
endometabolome	2898
pure substance	2443
Colon	2361
mixture	2293
cervical mucus	2164
blank	2088
Plasma	1992
Serum	1950
liver	1941
fruit	1896
Sweat	1774
Seed	1727
rosette leaf	1197
supragingival dental plaque	1156
Wine	1087
...	

File Format	Count of files
.raw	72826
.mzML	52556
.mzXML	10590
.d	3907
raw.zip	1262

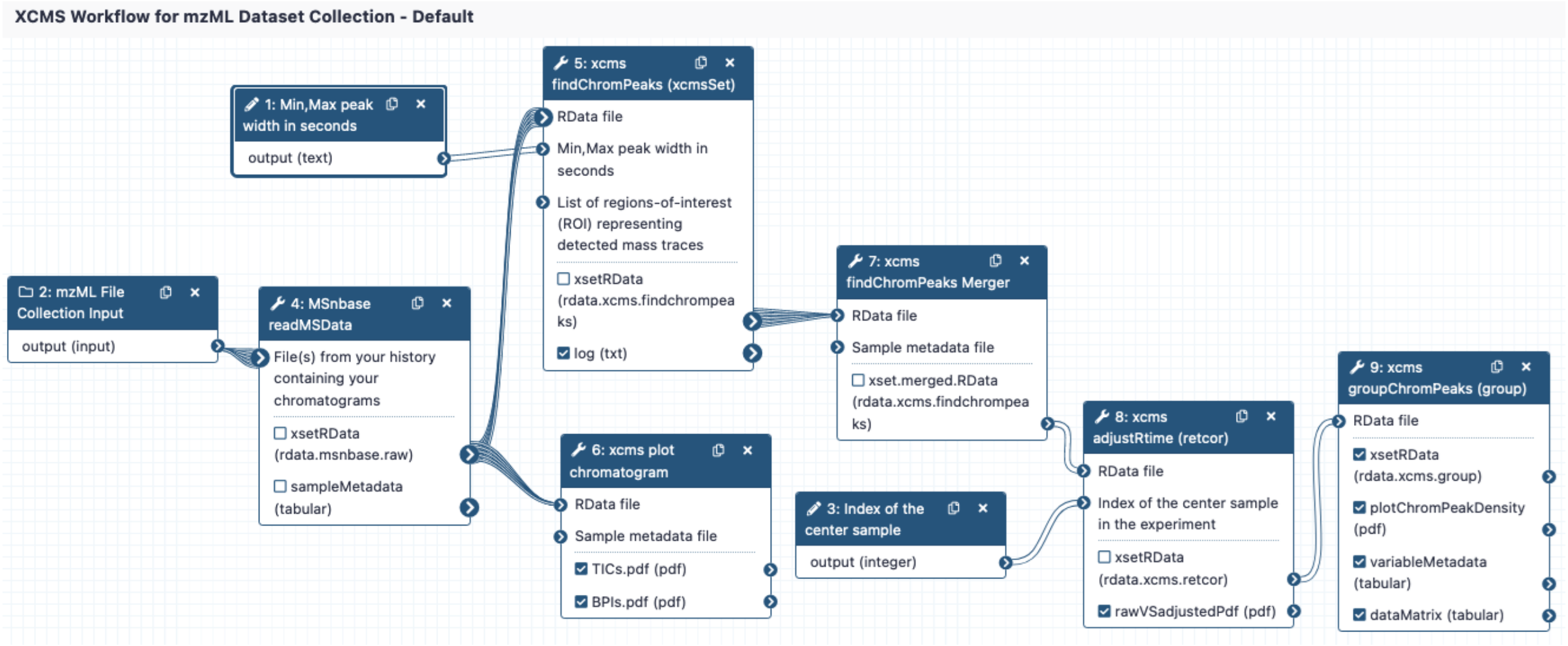
Galaxy Workflow: XCMS (& MSnbase)

- Individual XCMS steps are structured into a workflow to extract and match peaks across files and generate data matrix (m/z vs RT vs intensity)
- 4 parameter sets based on ‘assay type’ & ‘column model’
  - 50 cm length, UPLC
  - 100 cm length, UPLC
  - 150 cm length, UHPLC
  - 150 cm length, HPLC
- Optimisation in collaboration with Rick Dunn (UoL)



UNIVERSITY OF LIVERPOOL

Galaxy Workflow: XCMS (& MSnbase)



BEAMSpy

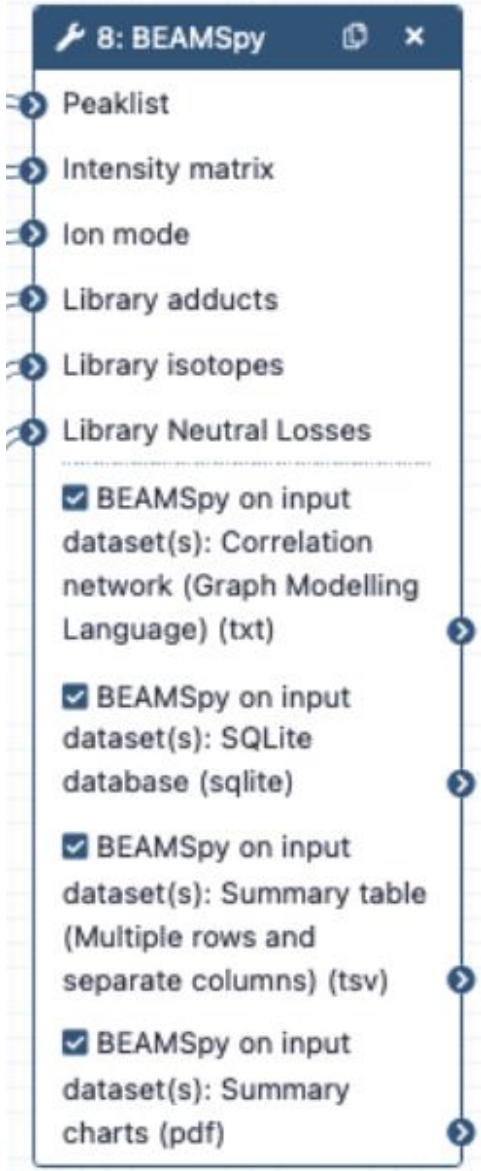
BEAMSpy is a Python package that includes **several automated and seamless computational modules** that are applied to putatively annotate metabolites detected in **untargeted ultra (high) performance** liquid chromatography-mass spectrometry or **untargeted direct infusion** mass spectrometry metabolomic assays.


README.rst

BEAMSpy - Birmingham mEtabolite Annotation for Mass Spectrometry (Python package)

pypi v1.2.0 | python 3.8 | 3.9 | 3.10 | repository GitHub | install with bioconda | beamspy passing | License GPL v3
documentation RTD | codecov 84% | launch binder

BEAMSpy (Birmingham mEtabolite Annotation for Mass Spectrometry) is a Python package that includes several automated and seamless computational modules that are applied to putatively annotate metabolites detected in untargeted ultra (high) performance liquid chromatography-mass spectrometry or untargeted direct infusion mass spectrometry metabolomic assays. All reported metabolites are annotated to level 2 or 3 of the Metabolomics Standards Initiative (MSI) reporting standards (Metabolomics. 2007 Sep; 3(3): 211–221. doi: 10.1007/s11306-007-0082-2). The package is highly flexible to suit the diversity of sample types studied and mass spectrometers applied in untargeted metabolomics studies. The user can use the standard reference files included in the package or can develop their own reference files.



Status Public | Release Date 2019-09-19 | Edit | ISA

MTBLS749: Alterations in the tyrosine and phenylalanine pathways revealed by biochemical profiling in cerebrospinal fluid of Huntington's disease subjects

Kim Kultima, Stephanie Herman

Huntington's disease (HD) is a severe neurological disease leading to psychiatric symptoms, motor impairment and cognitive decline. The disease is caused by a CAG expansion in the huntingtin (HTT) gene, but how this translates into the clinical phenotype of HD remains elusive. Using liquid chromatography mass spectrometry, we analyzed the metabolome of cerebrospinal fluid (CSF) from premanifest and manifest HD subjects as well as control subjects. Inter-group differences revealed that the tyrosine metabolism, including tyrosine, thyroxine, L-DOPA and dopamine, was significantly altered in manifest compared with premanifest HD. These metabolites demonstrated moderate to strong associations to measures of disease severity and symptoms. Thyroxine and dopamine also correlated with the five year risk of onset in premanifest HD subjects. The phenylalanine and the purine metabolisms were also significantly altered, but associated less to disease severity. Decreased levels of lumichrome were commonly found in mutated HTT carriers and the levels correlated with the five year risk of disease onset in premanifest carriers. These biochemical findings demonstrates that the CSF metabolome can be used to characterize molecular pathogenesis occurring in HD, which may be essential for future development of novel HD therapies.

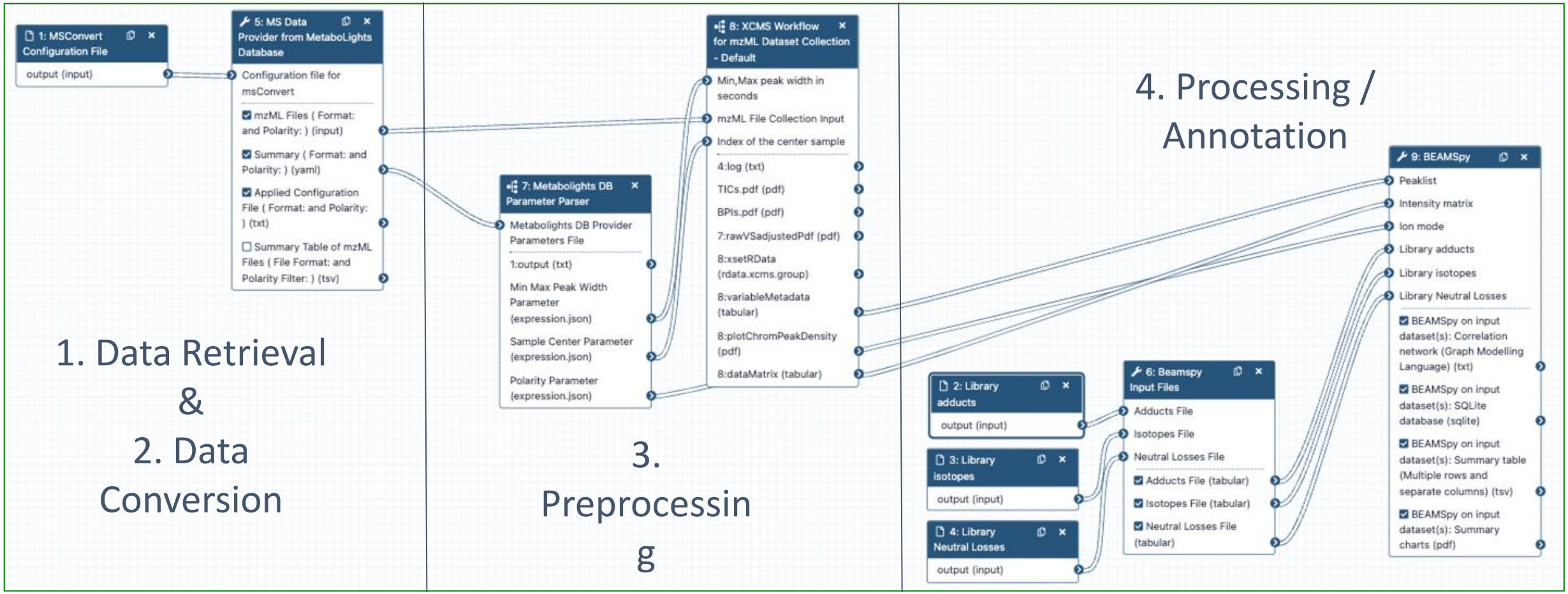
PUBLICATIONS

Alterations in the tyrosine and phenylalanine pathways revealed by biochemical profiling i...

Herman Stephanie, Niemelä Valter, Emami Khoonsari Payam, Sun...

Descriptors | Protocols | Samples | Assays | Metabolites | Files | Validations

MetaboLights Labs - Workflows - MDP + XCMS + BEAMSpy



Rick Dunn (PI),
Tim Ebbels (PI),
Claire O'Donovan (PI)

UNIVERSITY OF LIVERPOOL
Imperial College London
UNIVERSITY OF BIRMINGHAM
EMBL-EBI

In the Future

In the future we plan to:

- Integrate our workflows directly and seamlessly into submission process
- Develop workflows for the processing and annotation of MS/MS data, and facilitate additional analyses such as molecular networking
- Integrate MetaboLights and GNPS (as well as other repositories such as Metabolomics Workbench) to enable cross repository analysis
- Work with other communities to map standards (data and metadata) for example proteomics and SDRF (PRIDE)

PRIDE

PRoteomics IDEntifications Database

GNPS

Acknowledgements

MetaboLights
https://www.ebi.ac.uk/metabolights

EMBL-EBI Metabolomics team

- Claire O'Donovan (PI)
- Mark Williams
- Thomas Payne
- Noemi Tejera Hernandez
- Felix Amaladoss
- Callum Martin
- Ozgur Yurekten

UK Research and Innovation

EMBL

NIH

wellcome

elixir

Funded by the European Union

The Metabolomics team's activities were supported in the past year by the above funding bodies