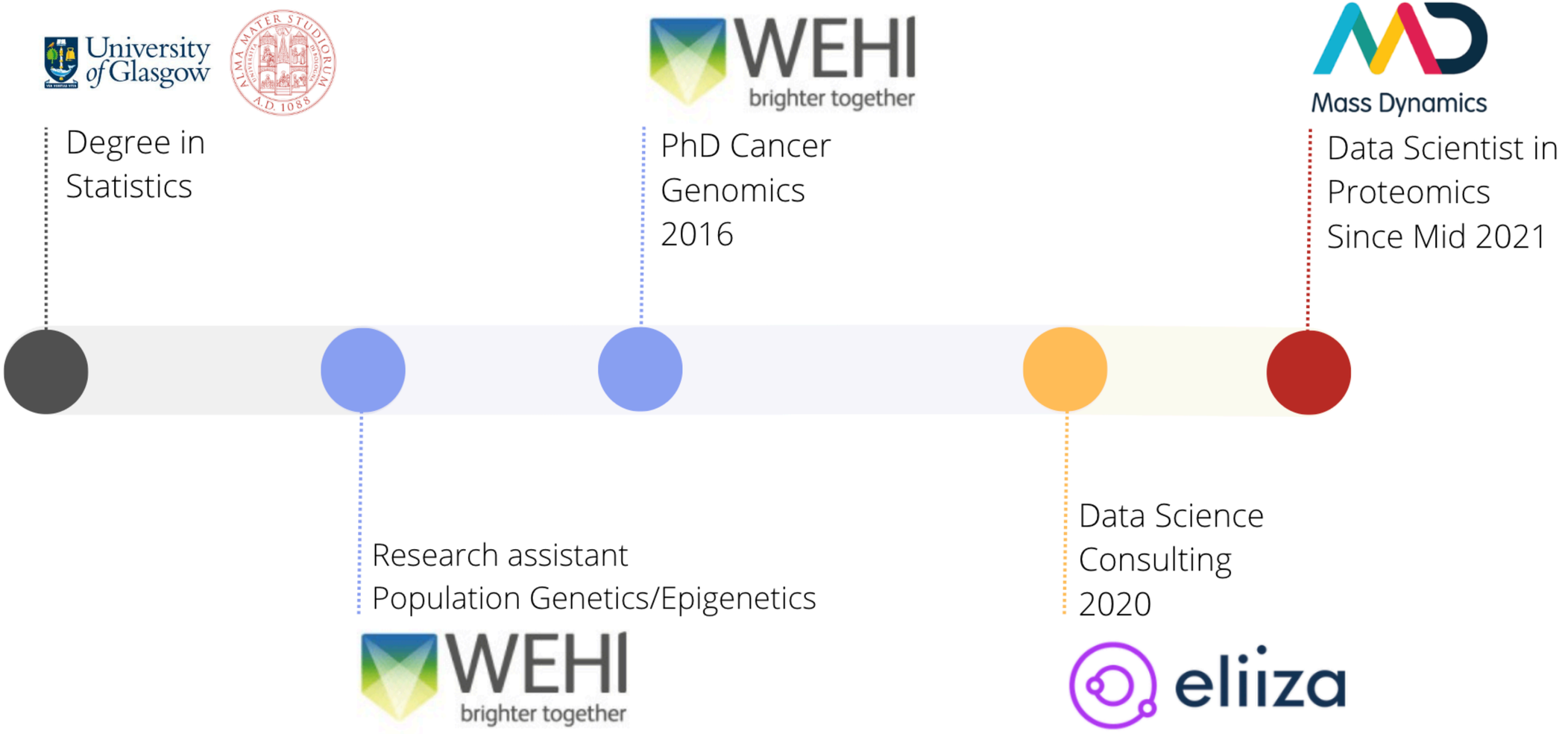
Introduction to Mass Spectrometry-based Proteomics data generation and processing

Anna Quaglieri 17/08/2022
Bioinformatics Data Scientist at Mass Dynamics

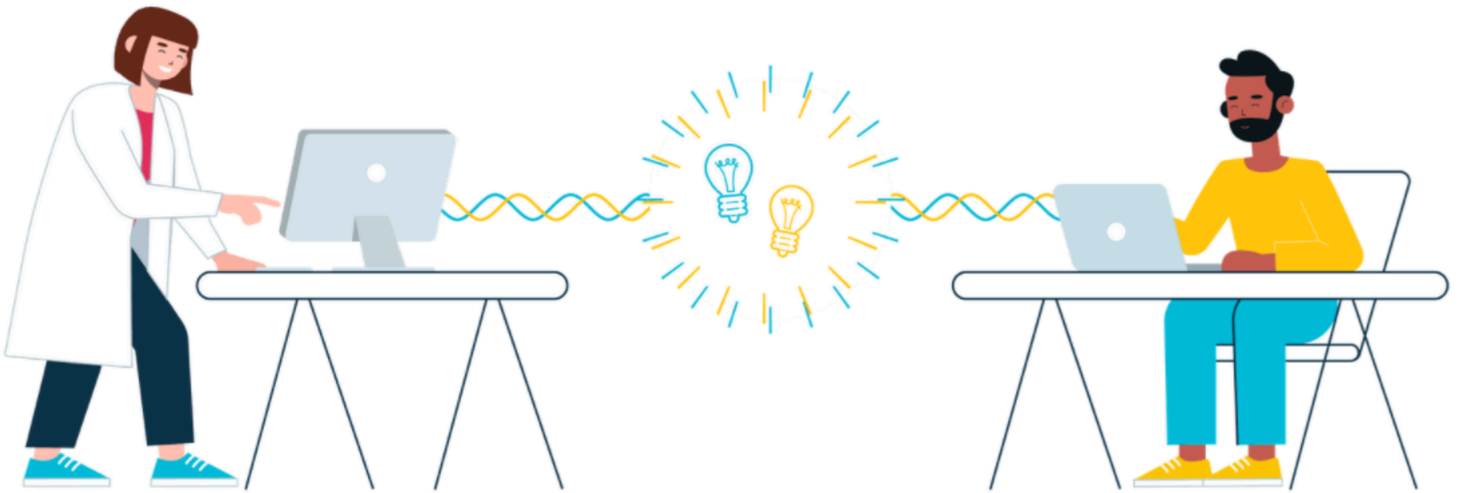
Mark Robinson laboratory
Department of Molecular Life Sciences, University of Zurich

About me


University of Glasgow
Degree in Statistics
WEHI brighter together
PhD Cancer Genomics 2016
Mass Dynamics
Data Scientist in Proteomics Since Mid 2021
Research assistant Population Genetics/Epigenetics
WEHI brighter together
Data Science Consulting 2020
eliiza

Mass Dynamics

Where life scientists discover, together.

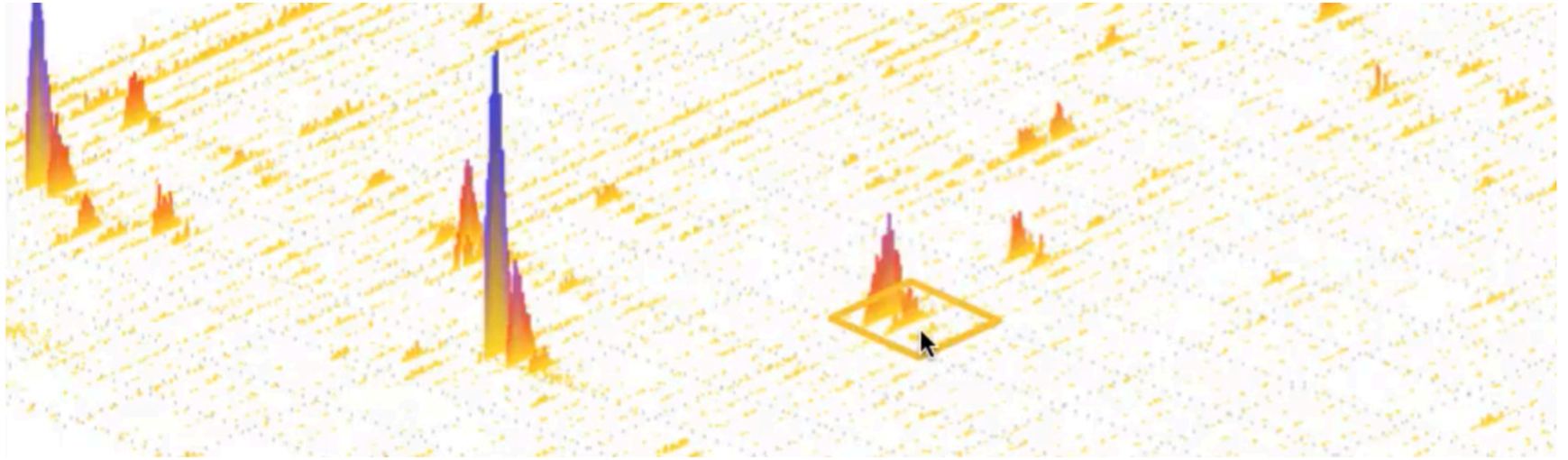


Mass Dynamics enables life scientists to transform complex and quantitative scientific data to knowledge - better, faster, easier and without compromise.

Mass Spectrometry data for proteomics

Outline

1. Mass Spectrometry (MS)-based proteomics data generation
    1. Data generation data
    2. Computational tasks in data processing
    3. Different technologies, different data
2. Common types of analyses
3. Commonly used methods and software
4. Useful links

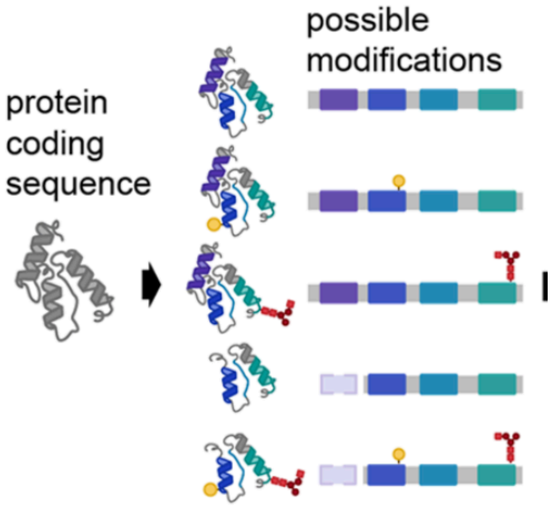


Mass Spectrometry based proteomics data generation

Proteomics vs Genomics

Proteomics

Large scale study of proteins: sequence, structure, abundance, modifications etc..

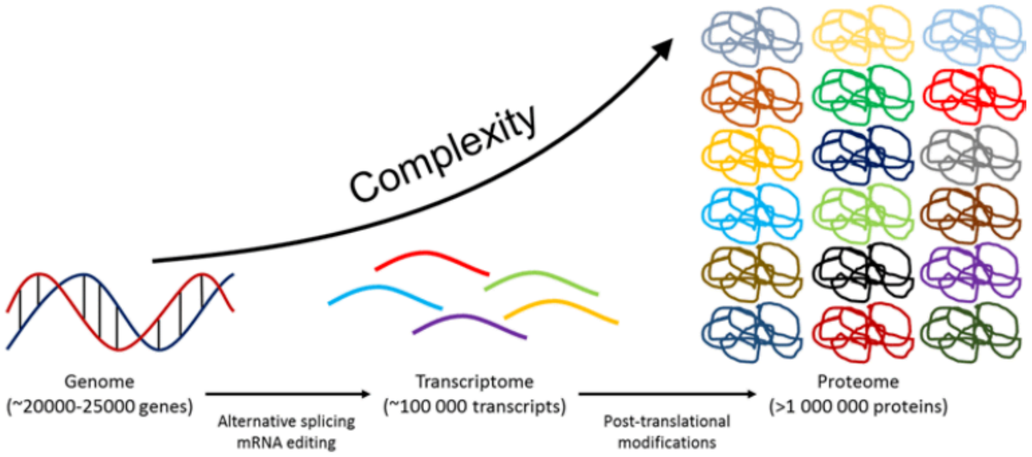


Proteomics is complex:

Human genome: ~20k genes

Human proteome can vary: ~50k - 500k proteins

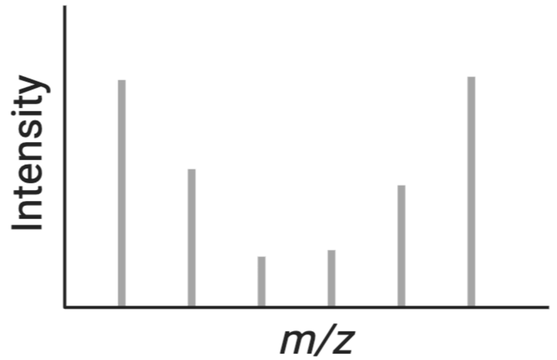
We cannot amplify proteins



Proteoforms

https://pubs.acs.org/doi/10.1021/acs.jproteome.1c00894
https://www.researchgate.net/publication/313535513_Microreactor_for_electrochemical_conversion_in_drug_screening_and_proteomics

Mass Spectrometry (MS)
We measure Mass Spectra



A technique used to measure the **mass-to-charge ratio (m/z)** of **ions**.

Ions: atoms or groups of atoms carrying one or more + or - electrical charges.

In MS-based proteomics we measure **peptide ions**: peptides the carry a particular electrical charge

MS is one one of the most used technique to quantify proteins in a sample.

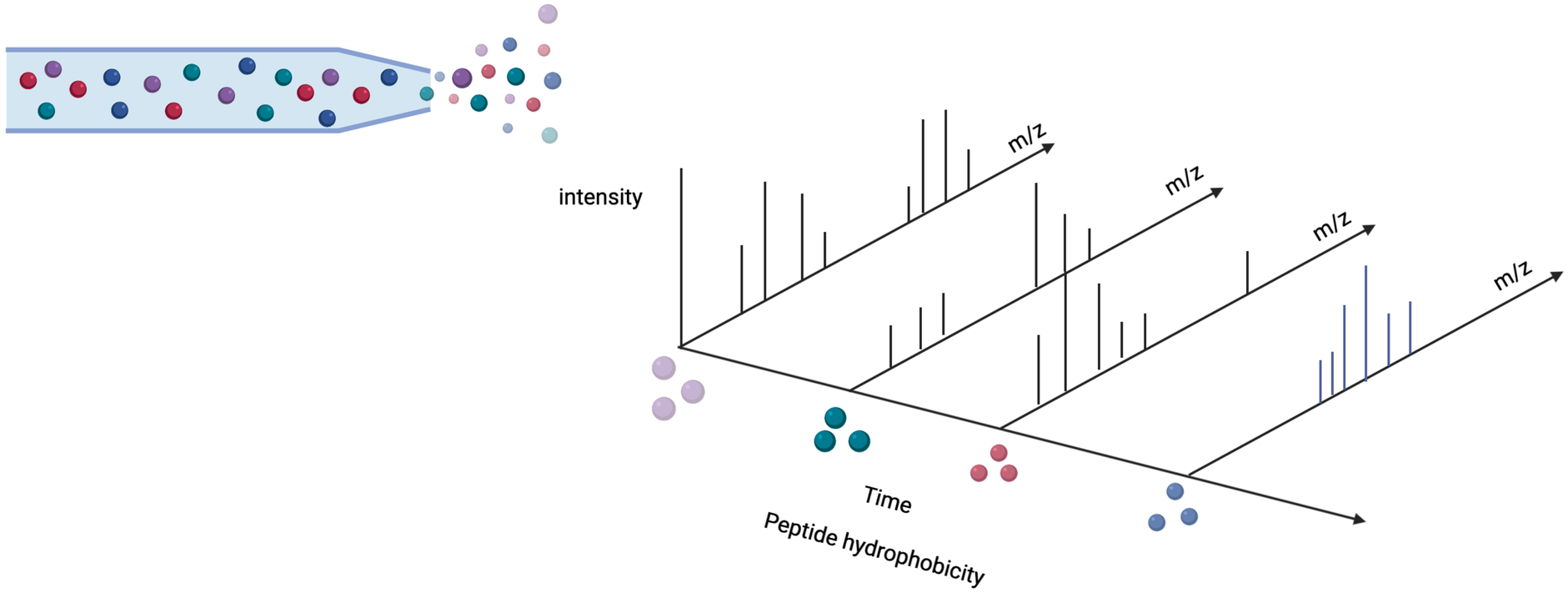
It is also used to quantify lipids and metabolites giving rise to broad large fields of research known as **Metabolomics** and **Lipidomics**.

Generated with Biorender

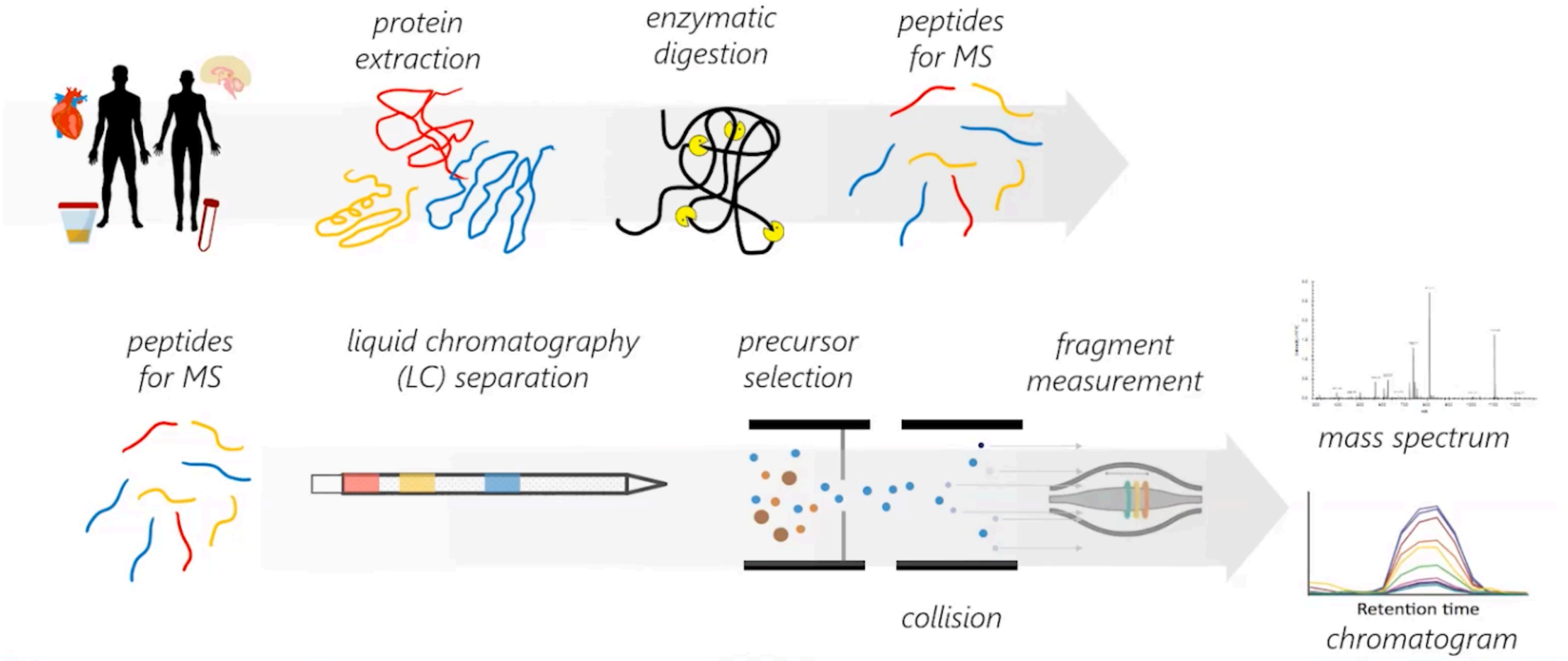
LC -MS/MS
Liquid Chromatography (LC) coupled with MS

MS is coupled with Liquid Chromatography to jointly detect and quantify peptides (and proteins).

LC exploits the hydrophobicity of peptides to separate them.



Overview of the mass spectrometry proteomics workflow



Lindsay K Pino, Talus Bio, USA

Introduction to Mass Spectrometry and Proteomics, HUPO: and https://vimeo.com/showcase/8948473/embed

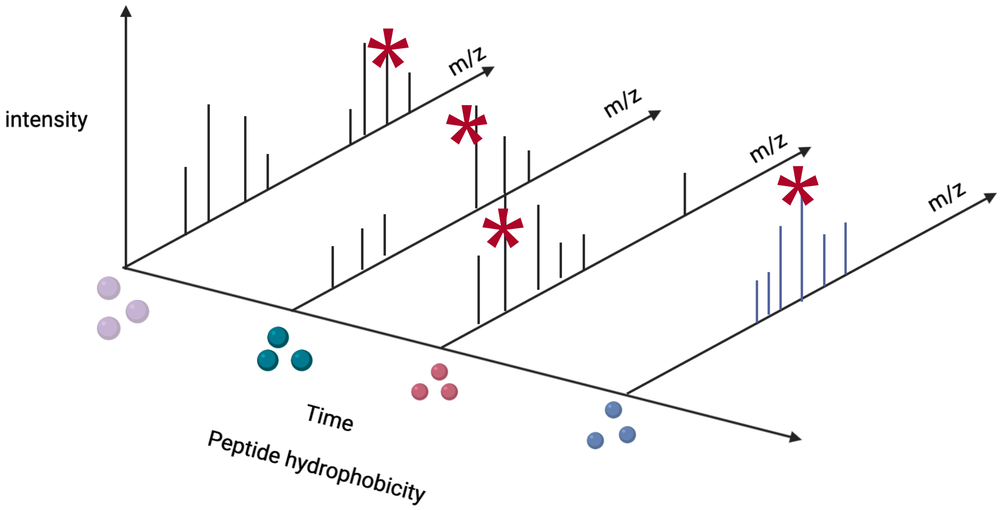
LC-MS/MS data

For each peptide ion going through the mass spec we get:

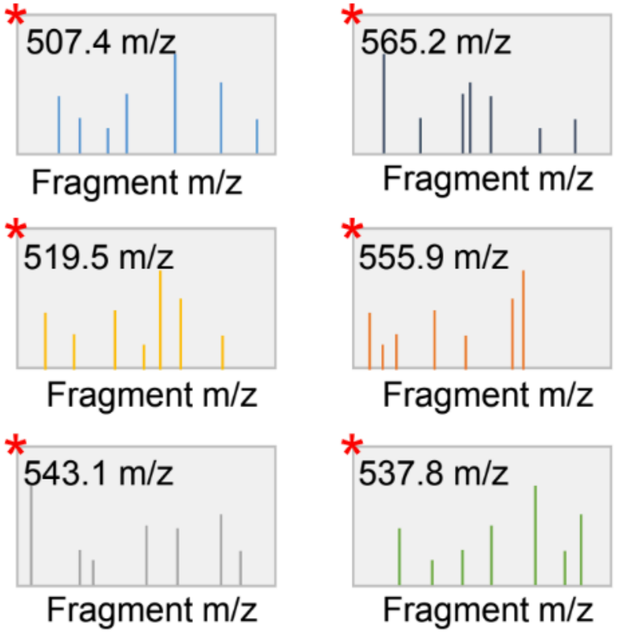
- m/z
- Retention time

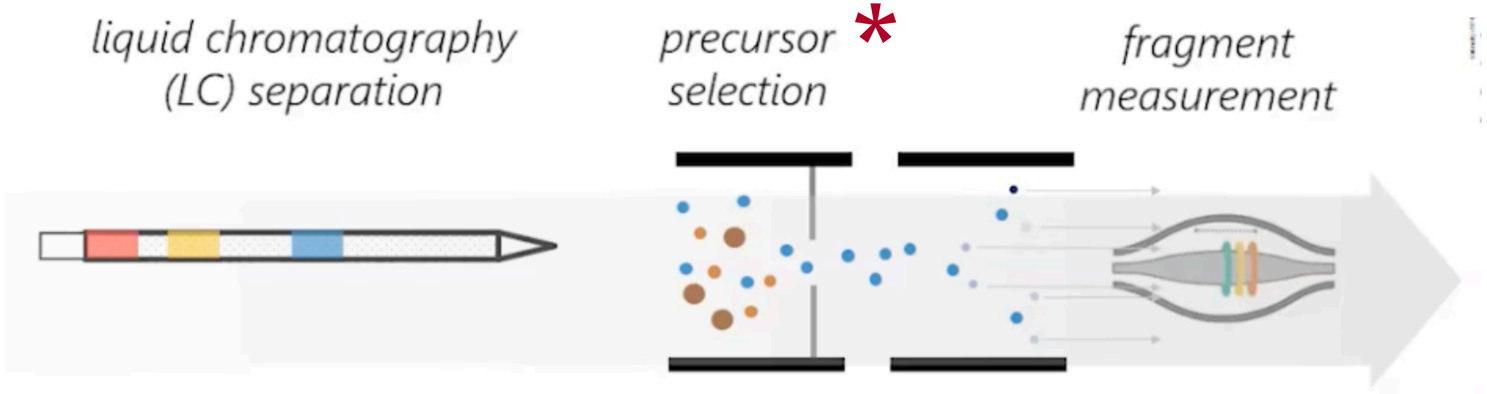
The y-axis intensity is a measure of how much biological material we found for that time and m/z

MS1 data



MS2 data





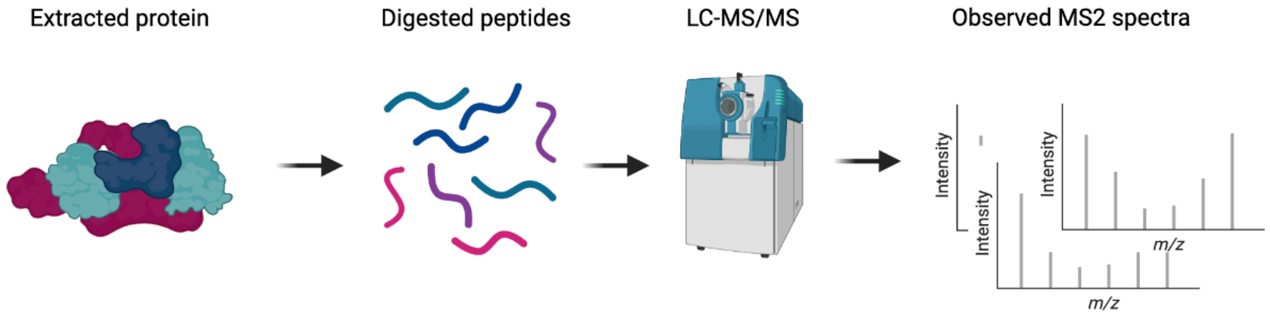
MS/MS (MS1/MS2)

Computational tasks: Identification

Peptide identifications: what peptides are found in a sample?

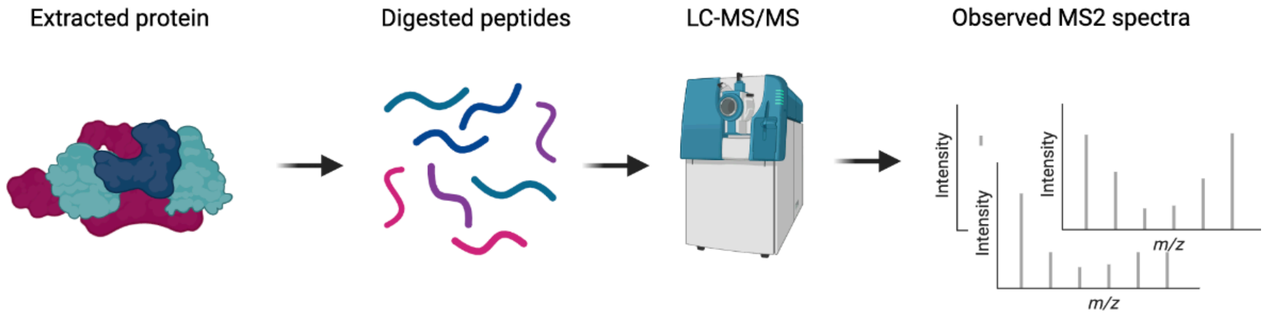
Computational tasks: Identification

Peptide identifications: what peptides are found in a sample?

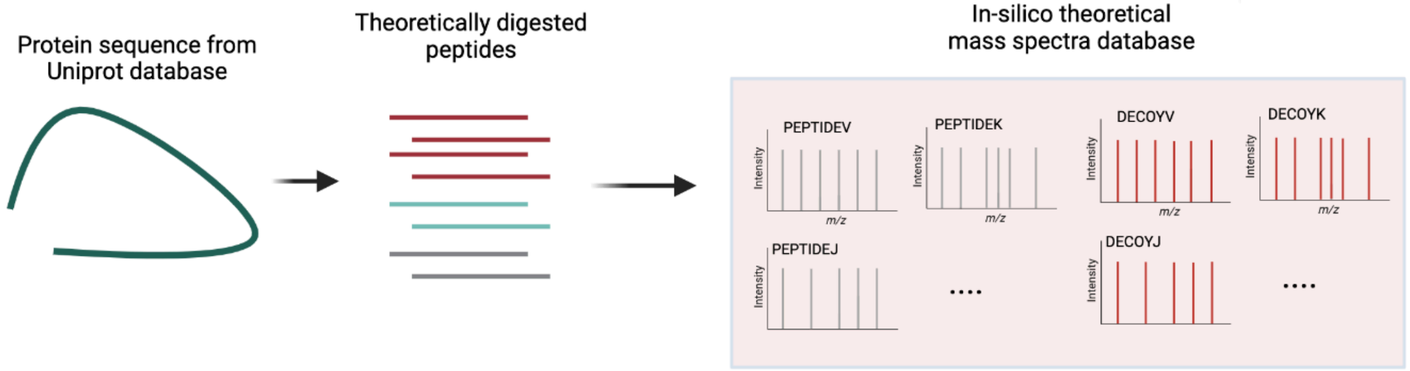


Computational tasks: Identification

Peptide identifications: what peptides are found in a sample?



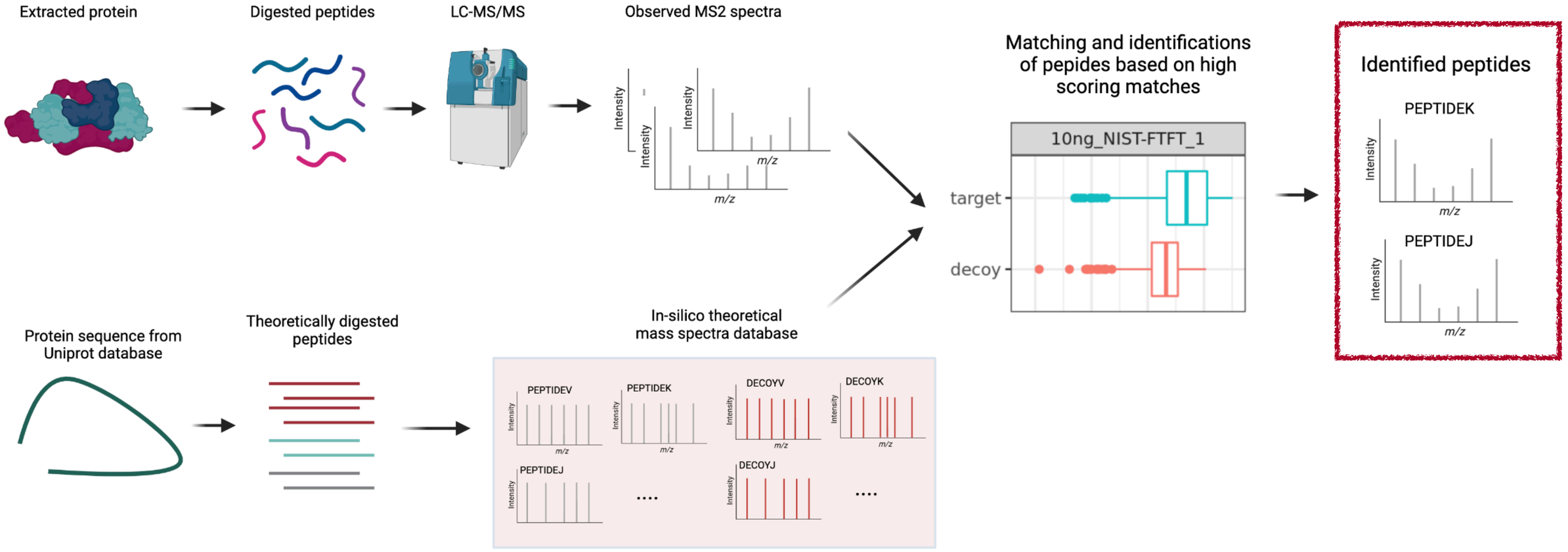
Observed spectra



Theoretical spectra

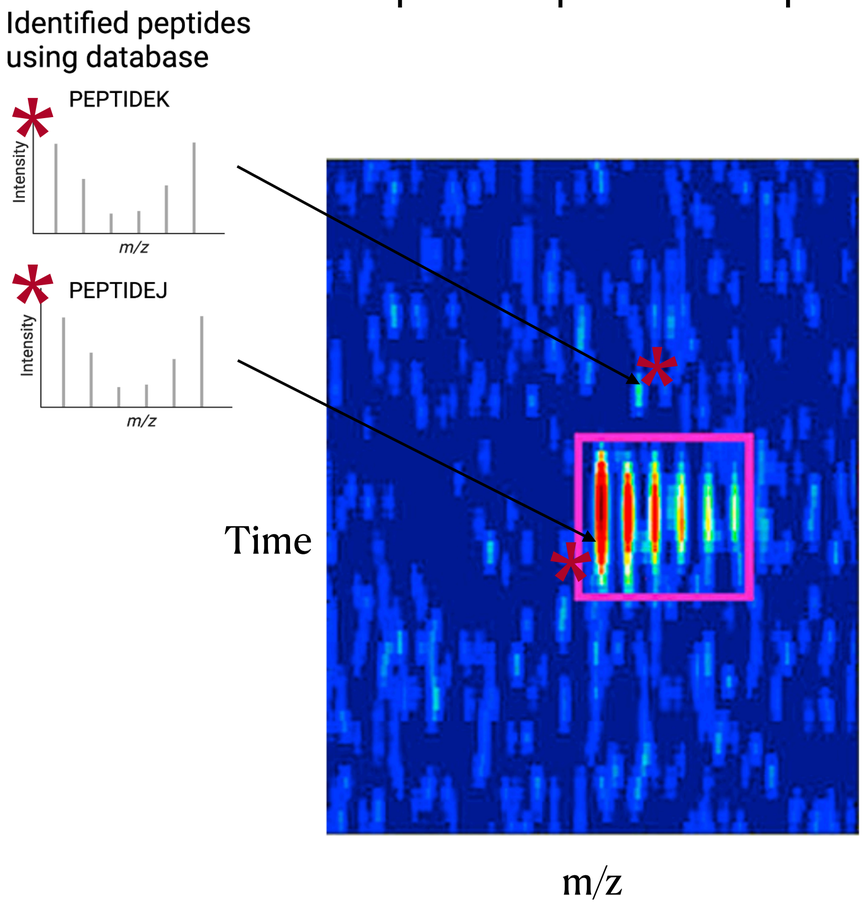
Computational tasks: Identification

Peptide identifications: what peptides are found in a sample?



Computational tasks: Quantification

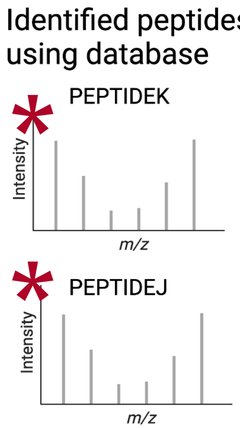
Peptide/protein quantification for downstream comparison across samples
Label Free Quantification

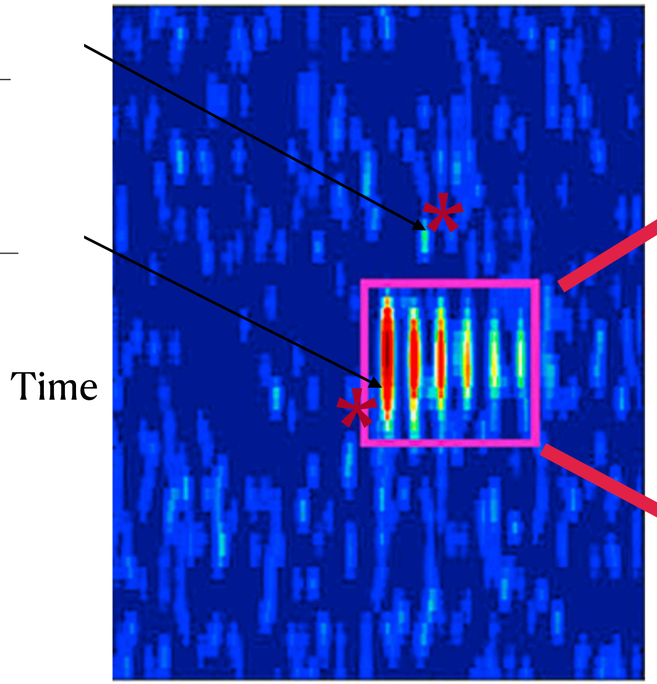


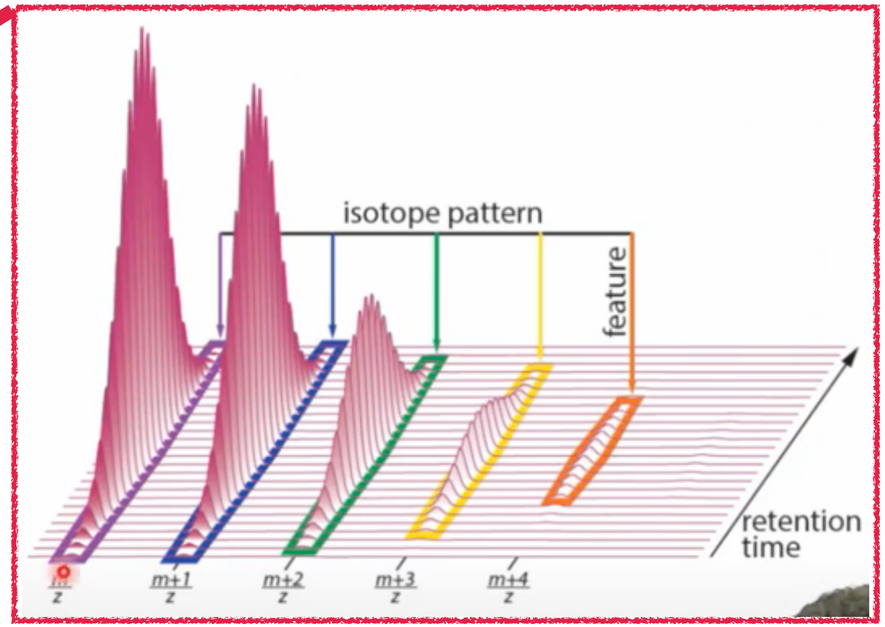
Heatmap from: https://www.sciencedirect.com/science/article/abs/pii/S0039914020308717#undfig1

Computational tasks: Quantification

Peptide/protein quantification for downstream comparison across samples
Label Free Quantification







Sum of the areas under the peaks
for all the isotopes of the peptide

Heatmap from: https://www.sciencedirect.com/science/article/abs/pii/S0039914020308717#undfig1

Final table of peptides/proteins intensities

Peptide Sequen	Sample1	Sample2	Sample3	Sample4
AGVVLKQ	12345	100000	0	0
WJKLGGFD	0	4545345	0	1000
TGFSSKHG	7568564	3523523	45678	1345
JKLKFSA	0	0	7888	8766
NNFDQKLHG		12578	0	0
FGDCBNJK	0	875645	1322	3432526

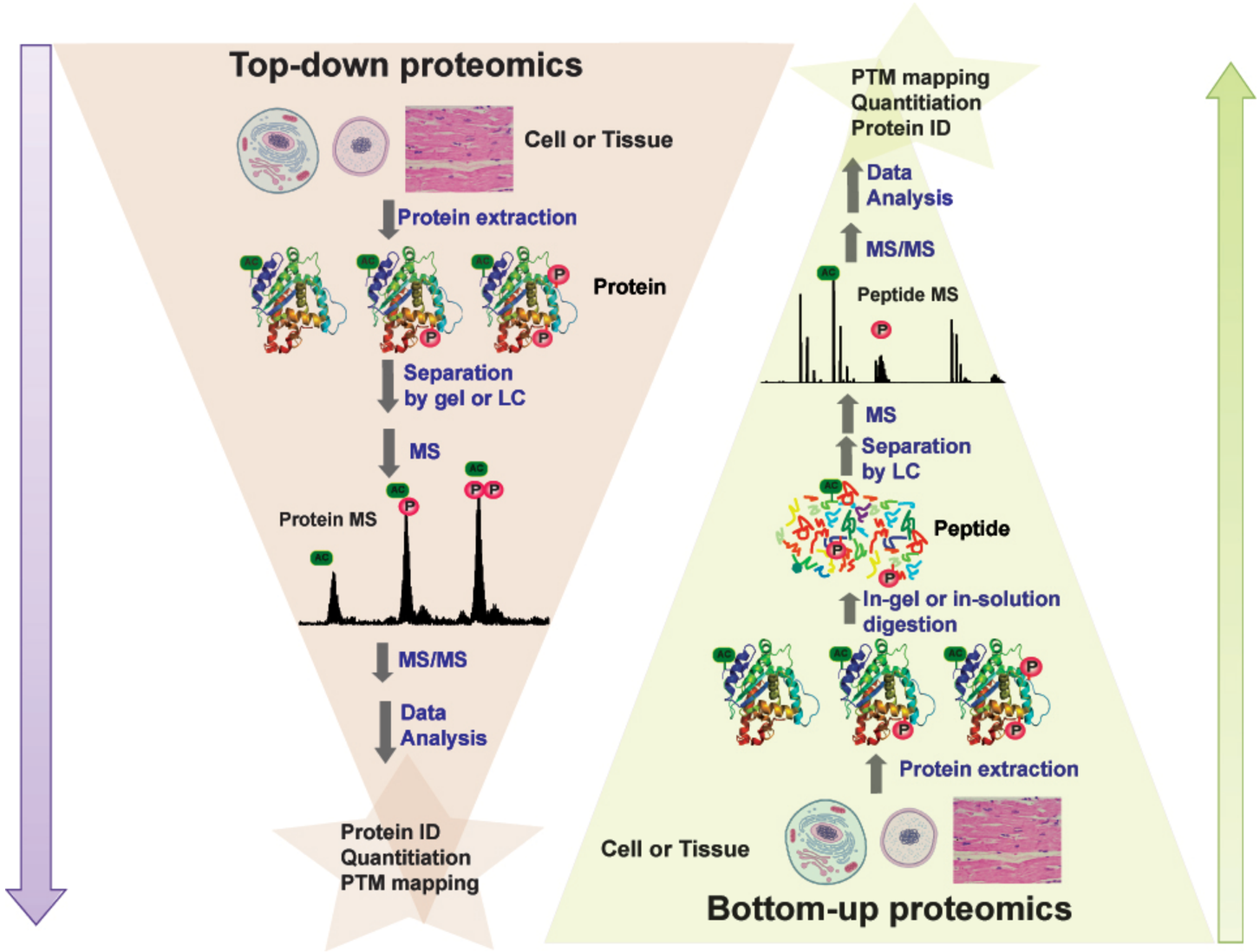
So far,

described the workflow for a *shotgun label free DDA*

LC-MS/MS experiment

Different LC-MS/MS technologies generate different data

Top-down vs Bottom-up (shotgun)



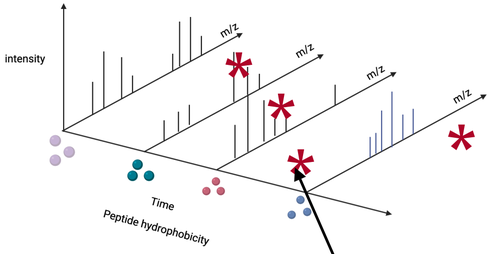
In Bottom-up Proteins are first digested (= divided) into peptides

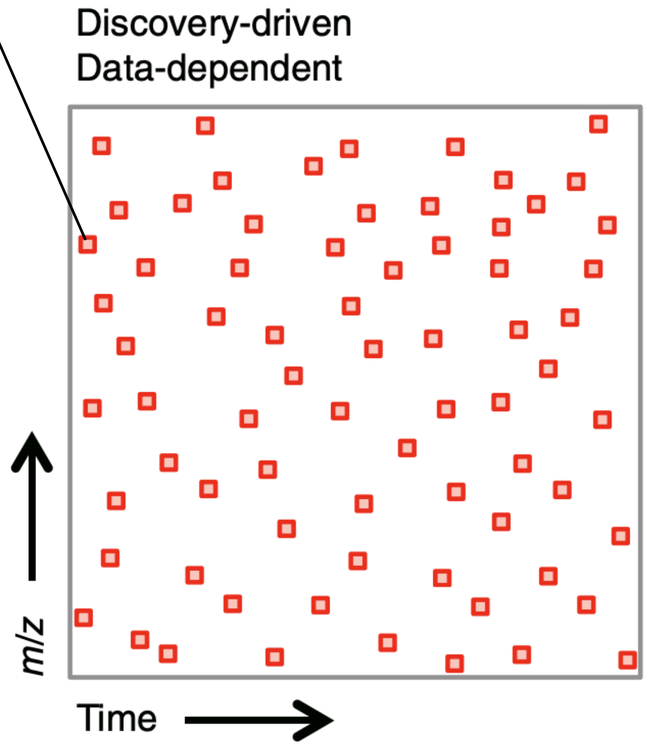
In Top-down Proteins are NOT digested but they are analysed directly:

- Harder
- Need more material
- Proteins can be really big and hard to handle

https://www.semanticscholar.org/paper/Proteomics-in-heart-failure%3A-top-down-or-bottom-up-Gregorich-Chang/26470f8be3017cb4f9b2a04ee1326457f668f252

Data acquisition



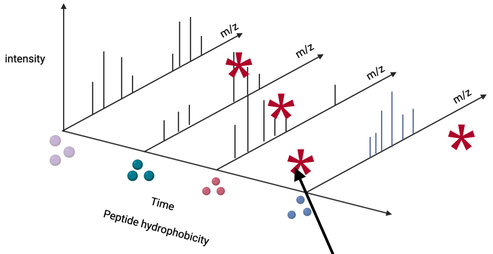


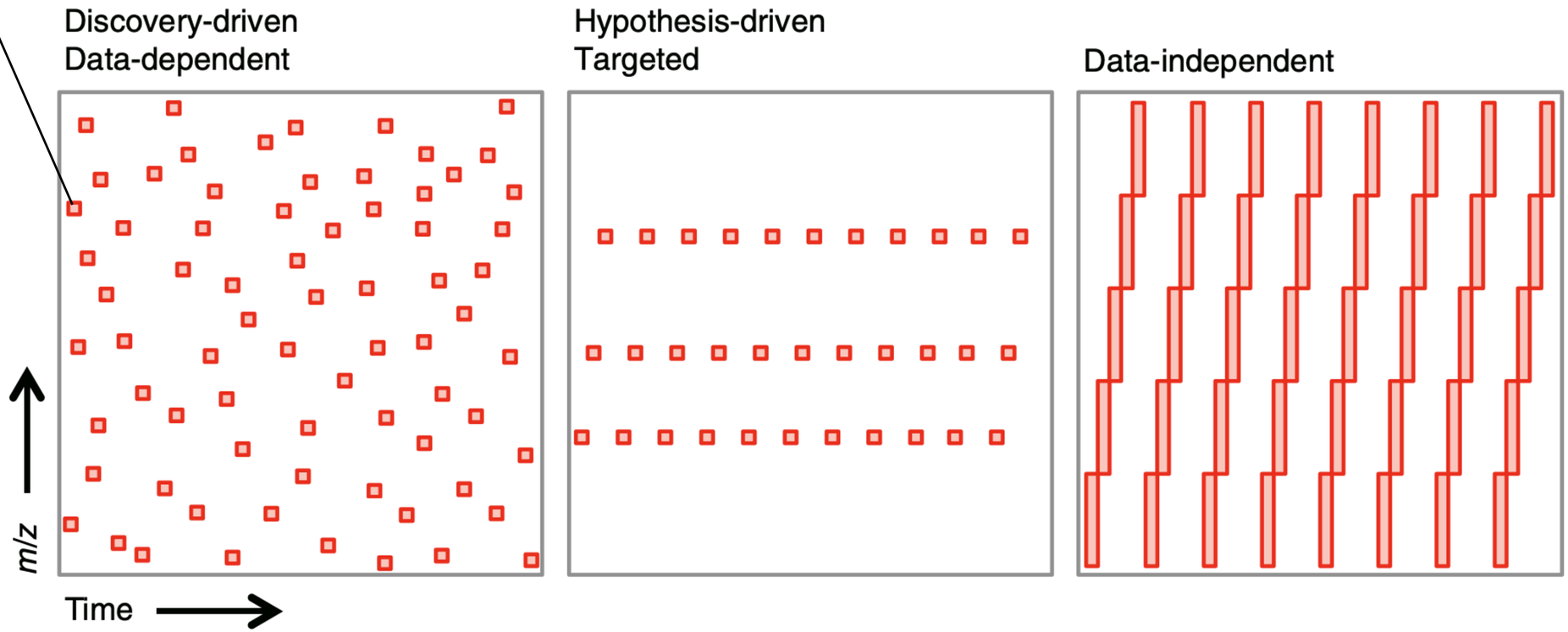
DDA

Stochastic selection dependent
on MS1 data
Missing values problem

Olga T Schubert et al, *Nature Protocols*, 2017

Data acquisition





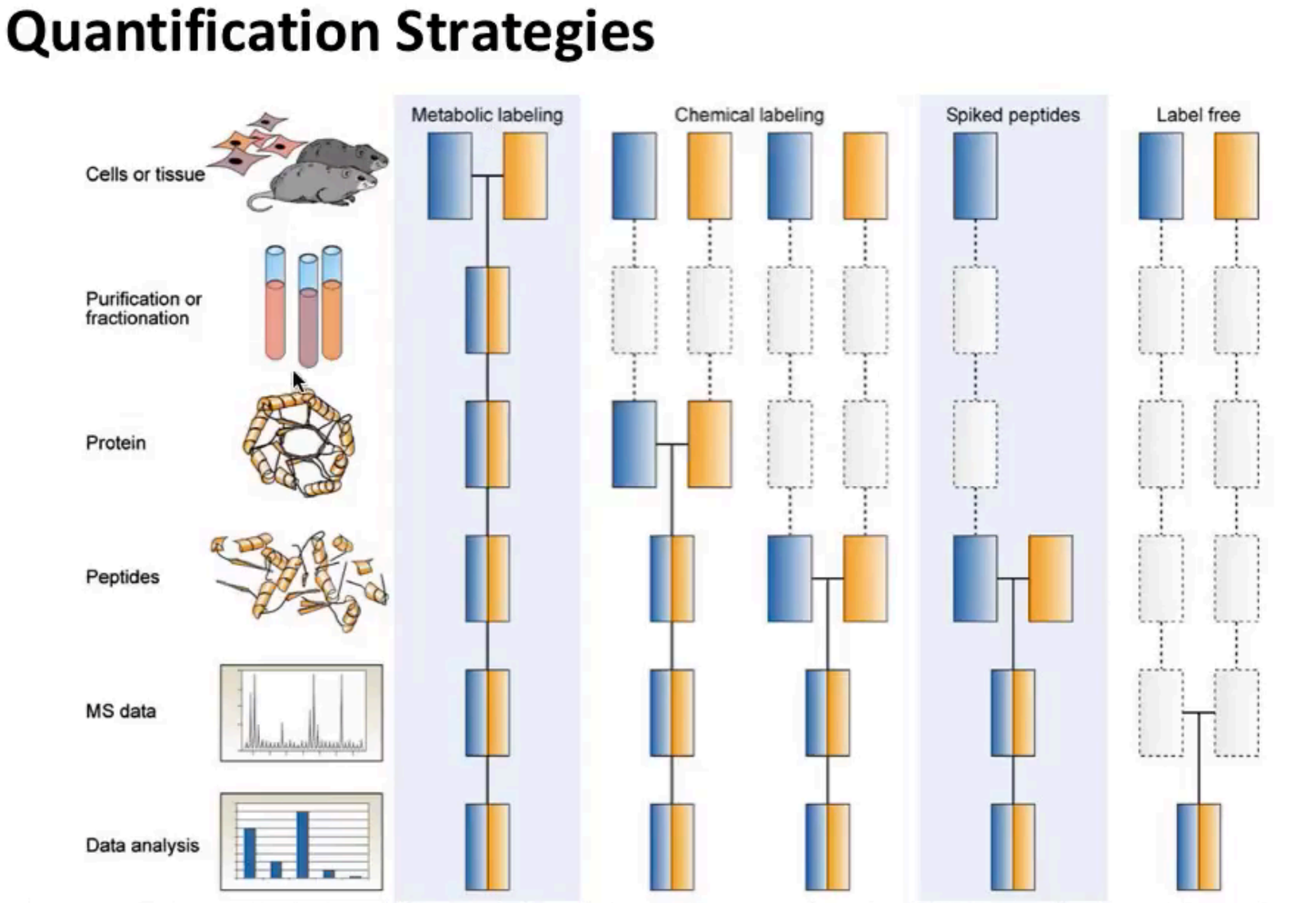
DDA
Stochastic selection dependent on MS1 data
Missing values problem

SRM/PRM
Deterministic selection of precursors based on prior knowledge

DIA
Deterministic selection of precursors for fragmentation

Olga T Schubert et al, *Nature Protocols*, 2017

Labelled vs Label Free Quantitative proteomics



Common quantitative mass spectrometry workflows. Boxes in blue and yellow represent two experimental conditions. Horizontal lines indicate when samples are combined. Dashed lines indicate points at which experimental variation and thus quantification errors can occur.

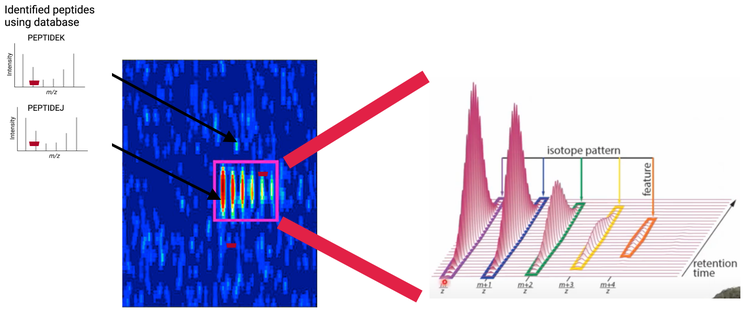
Bantscheff et al, Analytical and Bioanalytical Chemistry, 2007

Common types of analyses using MS-based proteomics data

Discovery type analysis

Identify and quantify peptides/proteins proteome wide for differential expression (abundance) analyses/ clustering/prediction

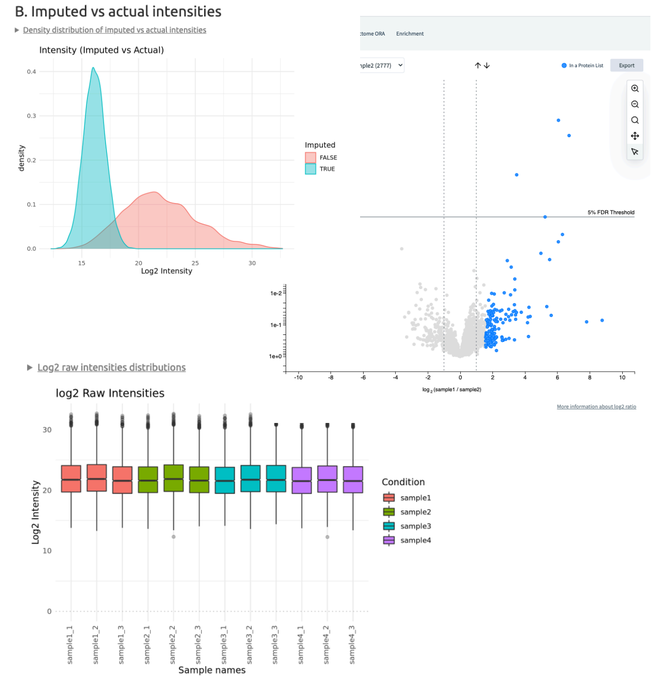
Pre-processing



Quantified data

Table of peptide intensities

Table of protein intensities

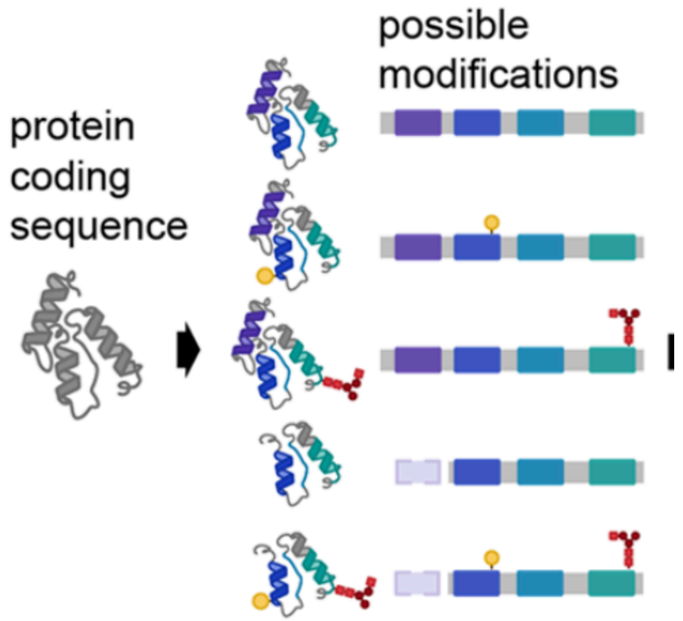


Downstream analysis of intensities
- Normalisation of intensities
- Imputation of missing values
- Differential expression
- Pathway analysis etc…

Post Translational Modifications (PTMs) identification

PTMs are chemical modifications of the proteins that can modify the activity of the protein and how it interacts with other molecules

Usually the first step involves enrichment for specifics PTMs to avoid missing low abundant ones.



Examples

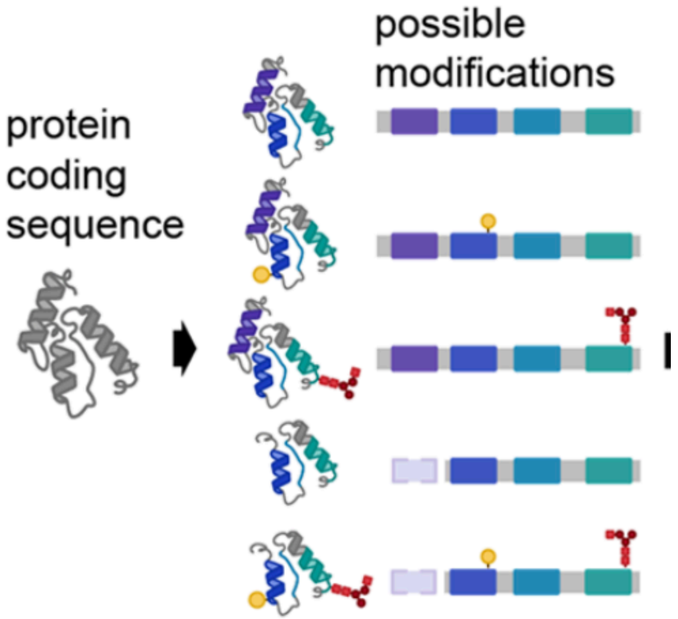
Methylation
Acetylation
Glycosilation
Phosphorylation
Ubiquitination

https://www.sciencedirect.com/science/article/abs/pii/S0021967322005453

Post Translational Modifications (PTMs) identification

PTMs are chemical modifications of the proteins that can modify the activity of the protein and how it interacts with other molecules

Usually the first step involves enrichment for specifics PTMs to avoid missing low abundant ones.



Examples

Methylation
Acetylation
Glycosilation
Phosphorylation
Ubiquitination

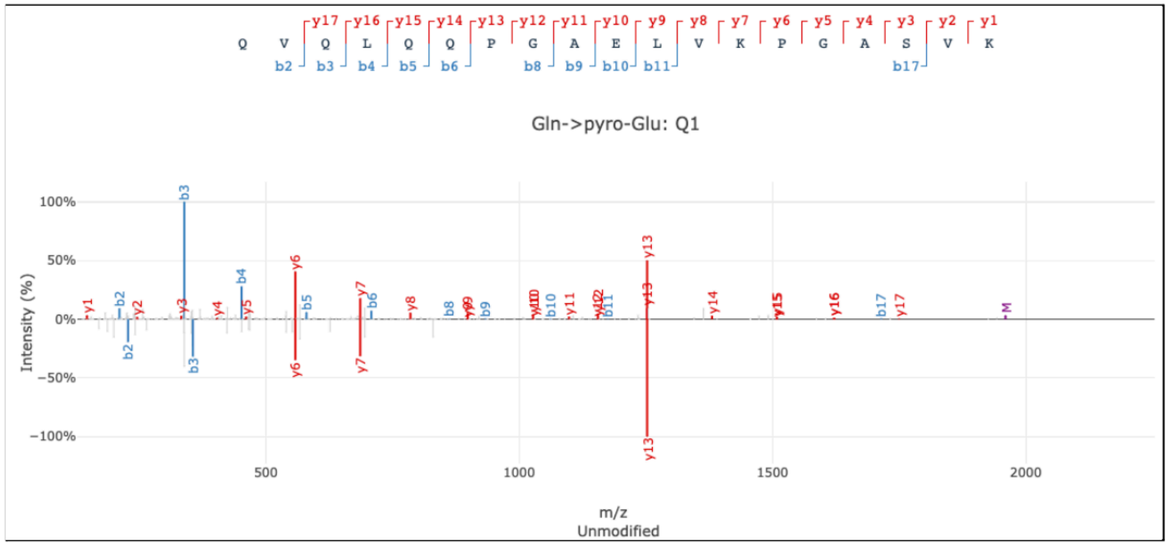
One of the most well-studied
It has its own large branch of proteomics known as phospoproteomics

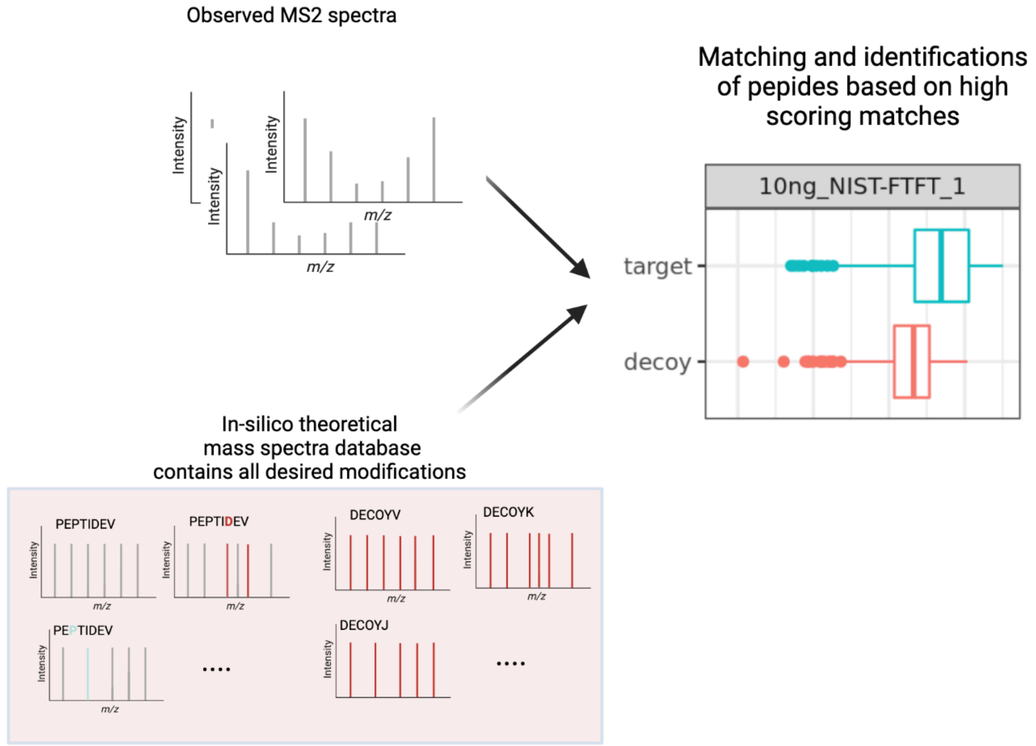
https://www.sciencedirect.com/science/article/abs/pii/S0021967322005453

Peptide mapping

Characterise and quantify post-translational modifications (PTMs) in target peptides/proteins.

Widely used for drug compounds characterisation.





Generated with Mass Dynamics

Generated with Biorender

What tools people use?

The choice of a software depends on

Your computational level expertise:

Is there a Mass Spec facility that can support you?

Do you have statistical/programming knowledge to DIY?

Is there someone in your lab that can help?

The type of data and analysis to be performed

Differential expression vs PTMs

The type of technology

DIA vs DDA vs Targeted

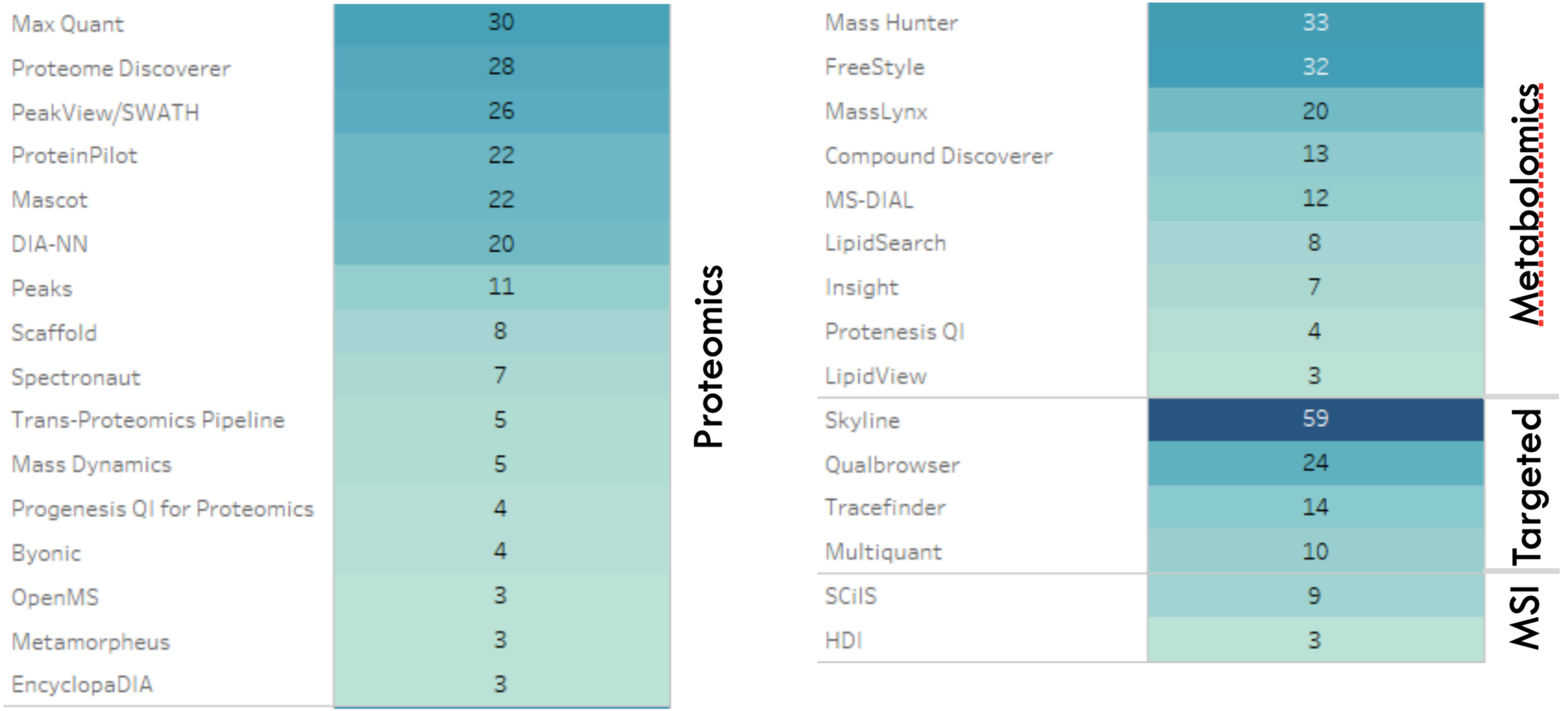
Label free vs Labelled

Australasia Core Facility Survey

- Results from the 3rd survey
- This round was open in July 2021, the majority of questions refer to the 2020 calendar year.
- 48 facilities invited, 37 responses received, inc. 6 from NZ.

THE UNIVERSITY OF SYDNEY

Software


Max Quant 30
Proteome Discoverer 28
PeakView/SWATH 26
ProteinPilot 22
Mascot 22
DIA-NN 20
Peaks 11
Scaffold 8
Spectronaut 7
Trans-Proteomics Pipeline 5
Mass Dynamics 5
Progenesis QI for Proteomics 4
Byonic 4
OpenMS 3
Metamorpheus 3
EncyclopaDIA 3
Proteomics
Mass Hunter 33
FreeStyle 32
MassLynx 20
Compound Discoverer 13
MS-DIAL 12
LipidSearch 8
Insight 7
Protenesis QI 4
LipidView 3
Skyline 59
Qualbrowser 24
Tracefinder 14
Multiquant 10
SCiIS 9
HDI 3
Metabolomics
Targeted
MSI

My breakdown of methods for data Processing

I divided the methods into 3 classes:

1. Software suite for several types of analyses. Aimed at researchers, .e.g MaxQuant-Perseus/Skyline/Mass Dynamics/LFQ Analyst
2. Do-it-yourself: assemble existing software to make your own workflow
3. Software targeted at BioPharma

Class 1: Software suite

Free, need download and only available for Windows (sometimes Linux)

DDA

- MaxQuant + Perseus: DDA Label Free Quantification data processing and analysis

- LFQ Analyst: downstream analysis of MaxQuant data: https://bioinformatics.erc.monash.edu/apps/LFQ-Analyst/

DIA

- Skyline (also open source): DIA/SWATH, Targeted:

- DIA-NN (also open source): DIA data - Free

Needs paid licence

- Spectronaut (from Byognosis Zürich): DIA

Class 1: Software suite - Several workflows

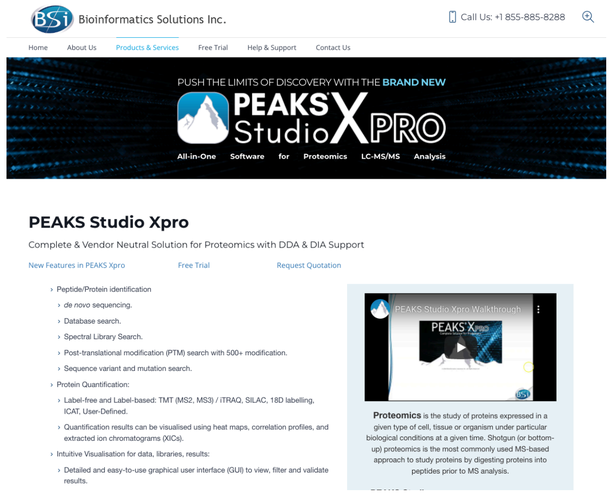
Free: Need download

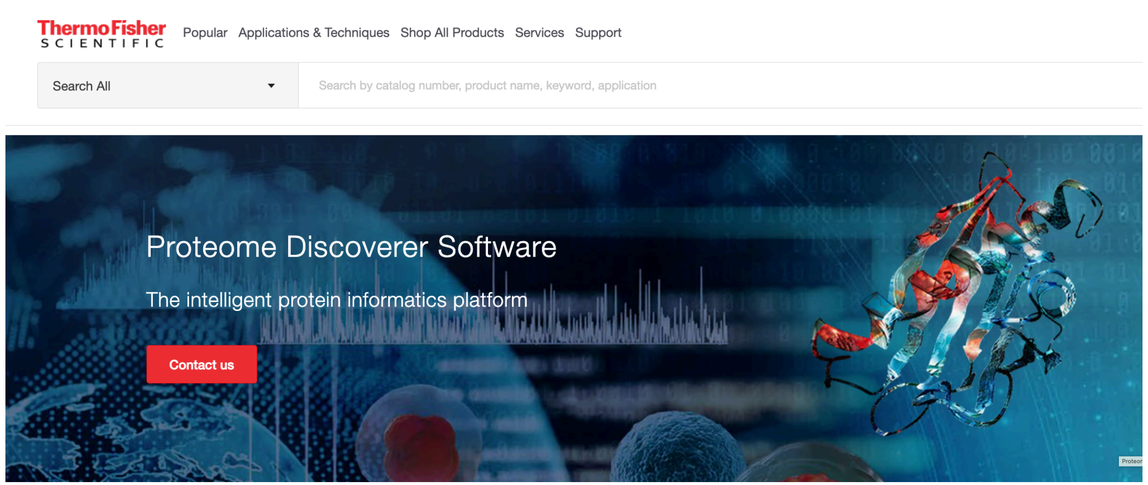
Progenesis QI (WATERS): several workflows supported

Licence $: Need download

- ProteomeDiscoverer: wide range of proteomics workflows

- Peaks





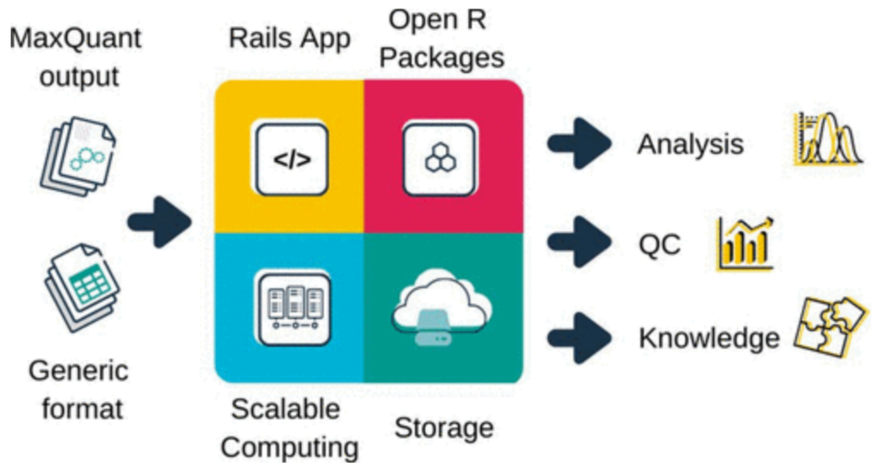
Class 1: Software suite - Cloud based

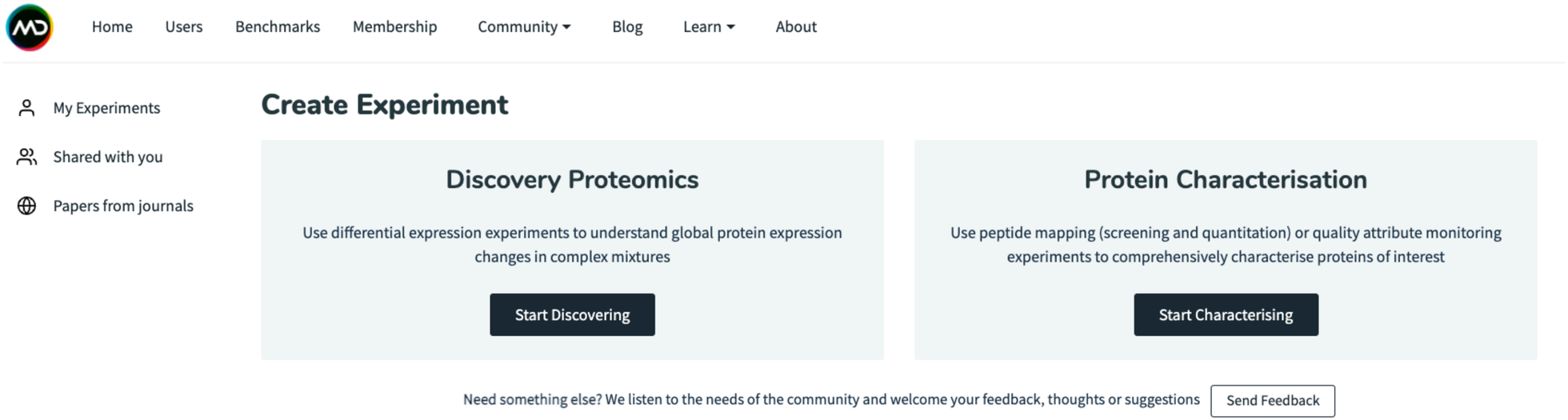
$ Membership or free

- Mass Dynamics (partially open source):

  downstream processing or from RAW (DDA data)

  DE, pathways analysis + Peptide Characterisation





Bloom J, et al, Journal of Proteome Research 2021

Class 2: Do-It-Yourself

For biologists/mass spec experts and developers with data analysis and coding skills

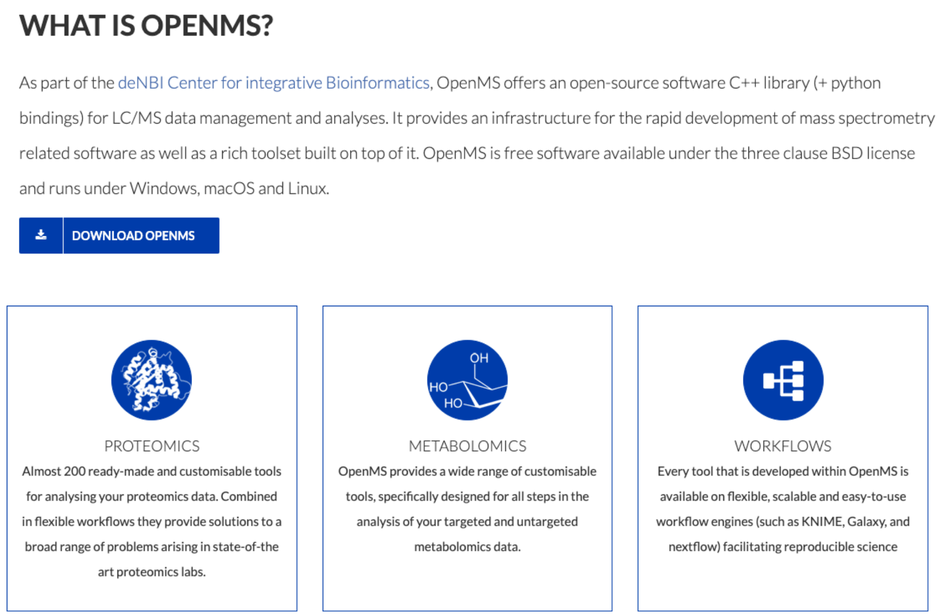
Most often open source libraries are used and allow flexibility and extensibility of new features

Class 2: Do-It-Yourself

OpenMS

https://www.openms.de/

Several tools for preprocessing of spectra , search, quantification



Full list of available tools

Class 2: Do-It-Yourself

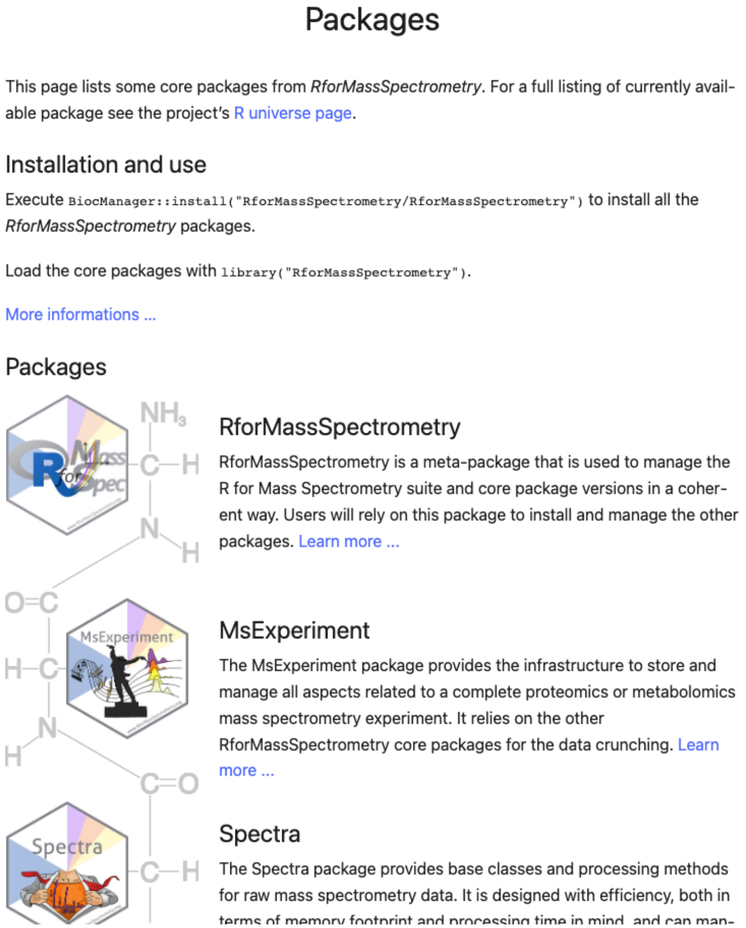
MSstats

Olga Vitek Lab (Northeastern University) R package introduction: https://www.youtube.com/watch?v=IYAqm2ENjkk&ab_channel=FenyoLab

Skyline provides an interface for MSstats

Class 2: Do-It-Yourself

R for Mass Spectrometry (https://www.rformassspectrometry.org/) & Bioconductor



Packages

This page lists some core packages from *RforMassSpectrometry*. For a full listing of currently available package see the project's R universe page.

Installation and use

Execute `BiocManager::install("RforMassSpectrometry/RforMassSpectrometry")` to install all the *RforMassSpectrometry* packages.

Load the core packages with `library("RforMassSpectrometry")`.

More informations ...

Packages

RforMassSpectrometry

RforMassSpectrometry is a meta-package that is used to manage the R for Mass Spectrometry suite and core package versions in a coherent way. Users will rely on this package to install and manage the other packages. Learn more ...

MsExperiment

The MsExperiment package provides the infrastructure to store and manage all aspects related to a complete proteomics or metabolomics mass spectrometry experiment. It relies on the other RforMassSpectrometry core packages for the data crunching. Learn more ...

Spectra

The Spectra package provides base classes and processing methods for raw mass spectrometry data. It is designed with efficiency, both in terms of memory footprint and processing time in mind, and can man-

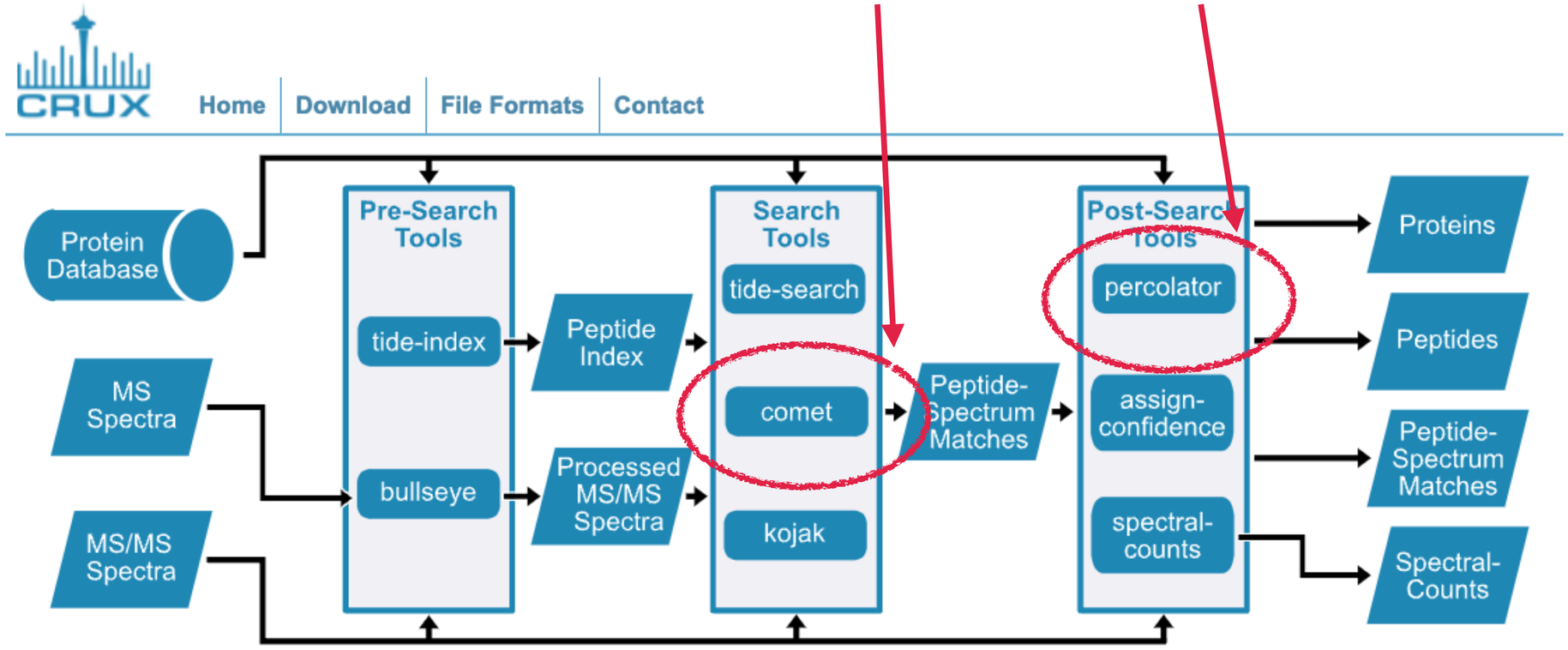
- Mostly useful to store information about experiments

- More useful to perform the downstream analysis once the RAW data have been pre-processed with MaxQuant, Skyline etc..

Class 2: Do-It-Yourself

Crux (~GATK for genomics)

Included in Mass Dynamics Peptide Mapping workflow



Example combining class 1 and 2

RAW input

Free
Desktop GUI for Windows

Label Free Quantification DDA data
All main labelling techniques DDA:
TMT, SILAC, iTRAQ

peptide/protein intensities

Statistical analysis:
Imputation, DE analysis, Pathway analysis etc..

Perseus (MaxQuant)
Free
Desktop GUI for Windows

Free R Shiny App (Monash Uni)

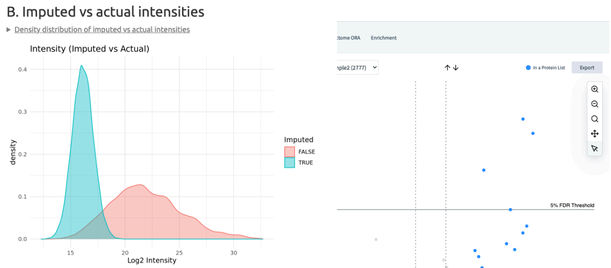
MSstats
Statistical Tool For Quantitative

Example combining class 1 and 2

RAW input
(LFQ DDA Thermo RAW)

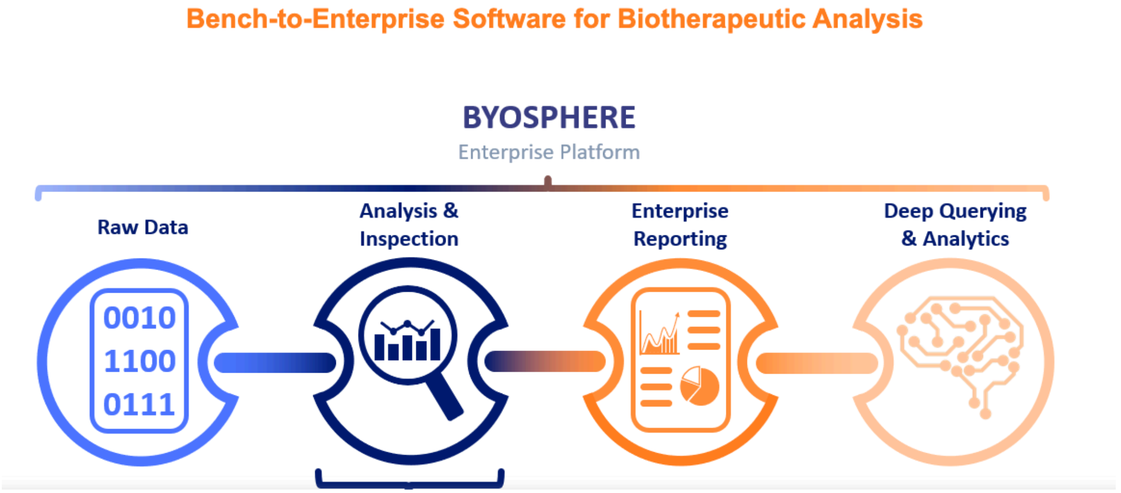
peptide/protein
Intensities input

Statistical analysis:
Imputation, DE analysis, Pathway analysis etc..



Class 3: For bio pharma

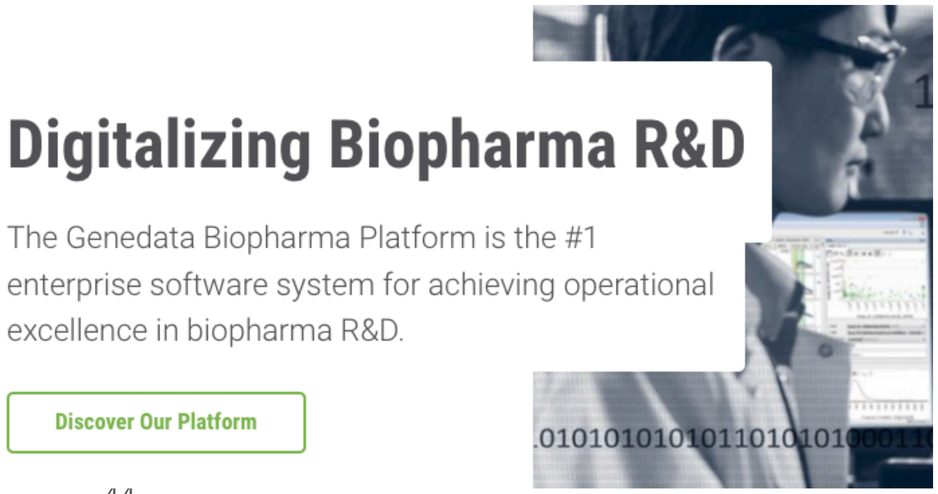
Software to support processing in BioPharmas



ProteinMetrics

Available for private cloud use

GeneData



Concluding remarks

The landscape of MS Proteomics data types is varied leading to different computational challenges

I've shared the most common software used to process most types of Proteomics data

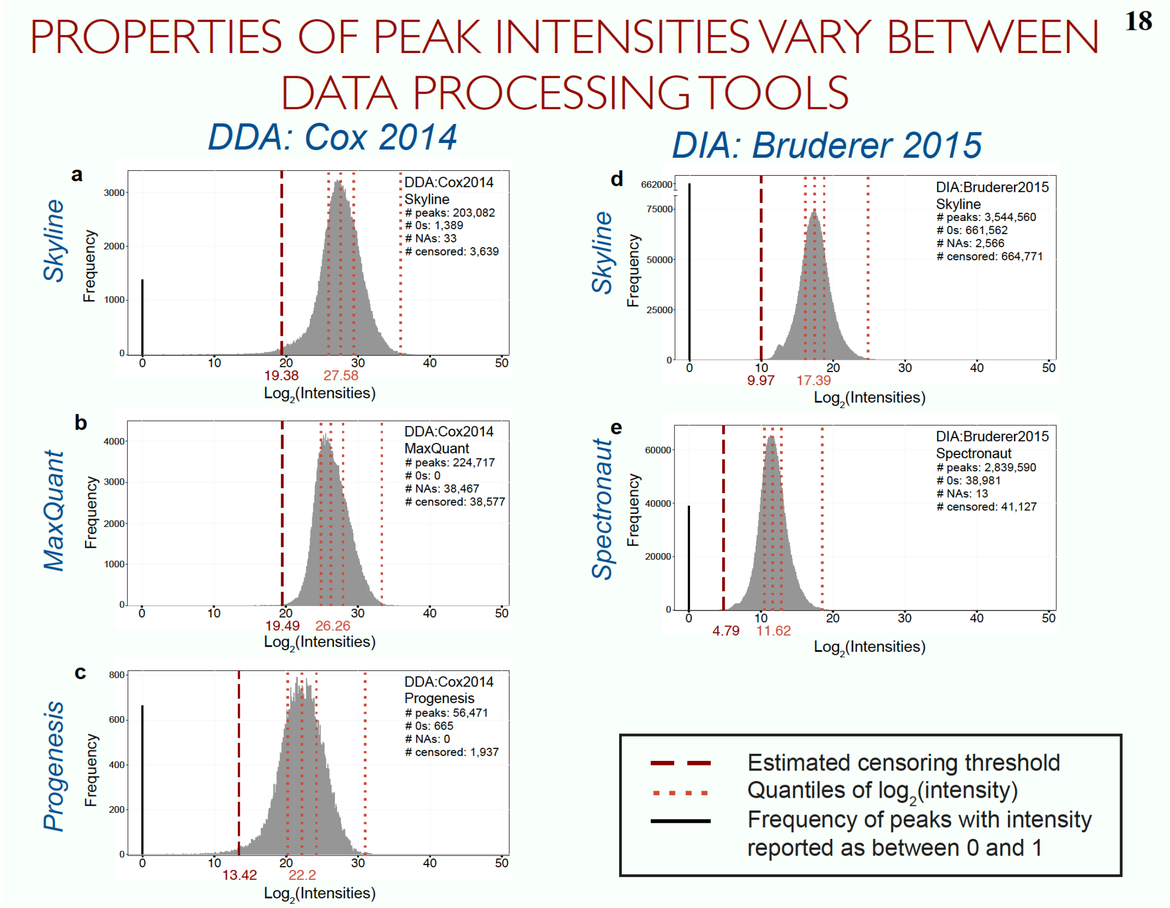
When starting your first ever MS Proteomics project as a computational person:

- Understand the type of analysis & experimental design required/used: e.g. DE analysis vs PTM analysis
- Understand the MS experiment design required/used: e.g. Acquisition type, Labelling, Fractionation etc..

I've noticed that it's not common in Proteomics to find big benchmarking comparisons to find what is the best methods

There are many methods, with several applications, all trying to improve something about the status quo

- Most choices would be ok, the main complications is in the quality of the MS data



Olga Vitek, MSstats tutorial, ASMS 2022

Workflow Benchmarking

Mass Dynamics regularly benchmarks all our workflows, to assess quality and measure performance as we continuously improve.

Peptide Mapping - *Probing the Sensitivity of the Lumos Mass Spectrometer using a Standard Reference Protein in a Complex Background**

The workflow is one of two product characterisation workflows. It is a targeted LFQ-DDA (Label Free Quantification-Data Dependent Acquisition) pipeline involving, feature detection, search, psm-scoring, ms1 recalibration and targeted feature detection (like match-between runs).

* "Probing the Sensitivity of the Lumos Mass Spectrometer using a Standard Reference Protein in a Complex Background"

Read

Regular benchmarks against ground truth to be aware of regression in the workflow performance

Useful links

Intro to MS and LFQ

Introduction to Mass Spectrometry and Proteomics from HUPO 2021: https://vimeo.com/showcase/8948473/embed

Olga Vitek: MSstats: an R package for quantitative MS-based proteomic experiments: https://www.youtube.com/watch?v=IYAqm2ENjkk

MaxQuant Basics 1, pre conference-course: https://www.youtube.com/watch?v=H_vClGghnNo

Label-Free quantitative proteomics: https://www.youtube.com/watch?v=p5cVkTuKWIQ

DIA-NN

High throughput proteomics with DIA-NN, Dr. Vadim Demichev, Single Cell Proteomics 2021

Single Cell Proteomics

Follow the Single Cell Proteomics conference YouTube channel, 2021/2022 conference introductions from Prof. Nikolai Slavov

Galaxy Training: Proteomics

Galaxy has loads of very useful training materials for how to get started using different software or how to perform different types of analyses.

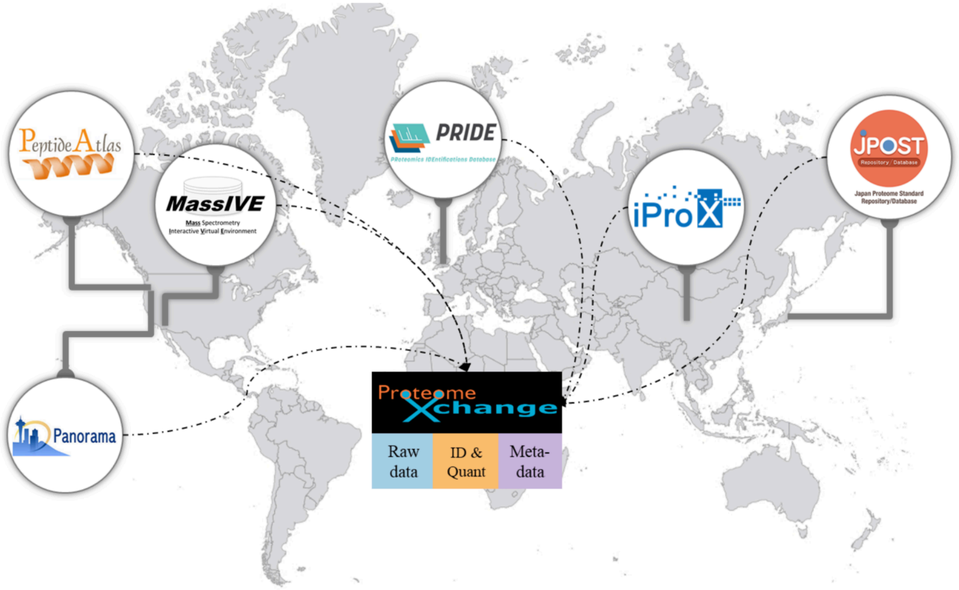
Pioneers in Zurich

Ruedi Aebersold, ETH Zurich, Biognosys was founded from his lab

Paola Picotti's lab, pioneer of LiP-MS technology

Proteomics data repositories and guidelines

ProteomeXchange

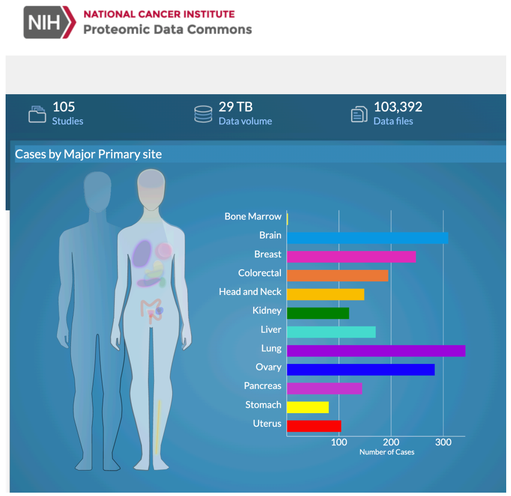


MassIVE.quant

Meena Choi et al, Nature Methods , 2020

Data Portal at CLINICAL PROTEOME TUMOR ANALYSIS CONSORTIUM (CPTAC)

Has genomics and Proteomics data



Acknowledgments

The source of all pictures and resources have been acknowledged in slides

ACF Survey summary acknowledgements

The Australasian Core Facilities Meetings are coordinated by Ben Crossett (USyd) and Ralf Schittenhelm (Monash).

This survey was originally designed with assistance from Paula Burton (Mass Dynamics), Mark Condina (while at UniSA) and the contents has been refined at the last two ACF meetings.

We would like to thank: Matt Padula (UTS), Tara Pukala (Adelaide) and Nick Williamson (UniMelb) for 'beta testing'; the 37 facilities that complete the survey and Naveed Nadiv (USyd) for expert assistance in compiling this summary.

Any enquiries, please contact: ben.crossett@sydney.edu.au