HBIGS course 2021/22: Introduction to computational RNA biology

Part 4: Proteomics and Integrative Analysis

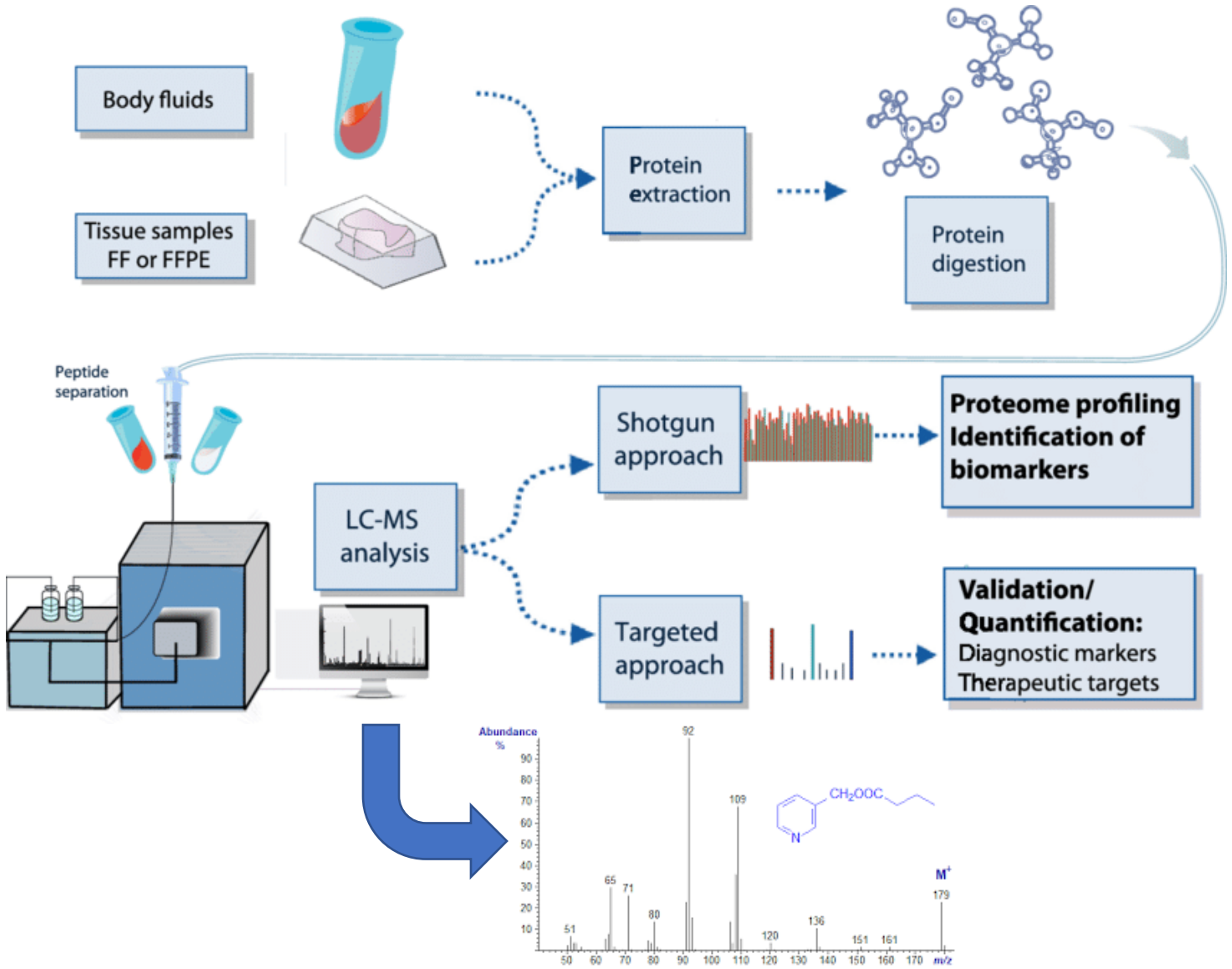
Enio Gjerga

Quantification of Protein Abundances

Quantification of Protein Abundances

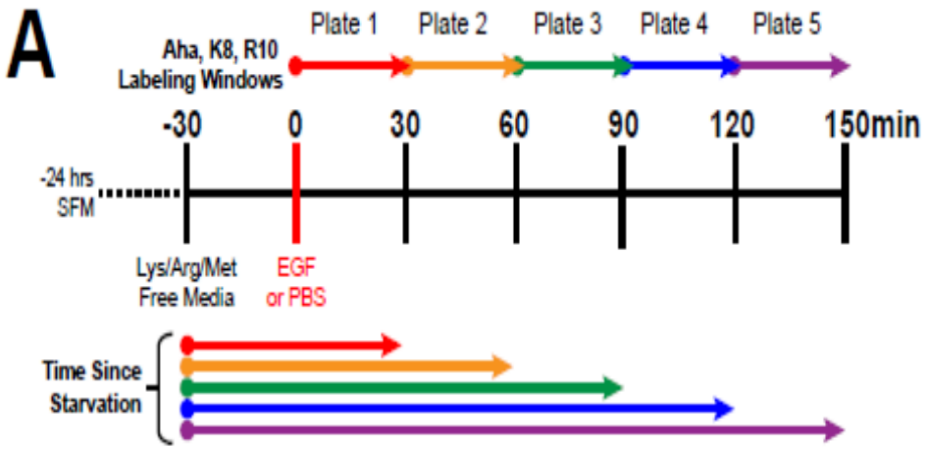
- Proteomics is a high throughput way to measure all proteins in a cell tissue by using mass-spectrometry.
- The goal of proteomics is to provide a time- and space- resolved description of a system's protein ensemble.
- Mass spectrometers are able to accurately determine the mass (more precisely, the mass-to-charge ratio, m/z) of biological molecules.
- There exist various types of mass-spectrometry based protein quantification techniques.

Mass-Spectrometry based Proteomics - Workflow



- Mass spectrometry can only measure over a limited mass range. Therefore, proteins are usually digested to peptides.
- The most common MS types introduce the compound into an HPLC (high performance liquid chromatography) analyzer. This requires for the peptides to be separated depending on their hydrophobicity (ensuring the ensure retention of hydrophilic proteins and elution of hydrophobic ones).
- Then the compound is ionized before entering the analyzer and then accelerated by a high electric voltage through a vacuum chamber before hitting the detector.
- For each passing peptide we obtain a spectra of m/z over abundance peaks which are compared to a peptide library.

Revealing EGF-driven protein synthesis across time



- MS was used to quantify proteome-wide protein synthesis rates temporally distinct time windows following EGF stimulation.

- HeLa cells were serum starved for 24 hr and stimulated with 20 nM EGF in 30-min windows, resulting in time points collected at 30, 60, 90, 120, and 150 min across 4 replicates.

- Matched PBS controls were also collected at the same time points.

Analysis of Proteomics Data-Sets

DEP R-Package for Proteomics Data Analysis

- **DEP** provides an integrated analysis workflow for robust and reproducible analysis of mass spectrometry proteomics data for differential protein expression or differential enrichment.

- It requires tabular input (e.g. txt files) as generated by quantitative analysis softwares of raw mass spectrometry data, such as MaxQuant or IsobarQuant.

- It includes tools to check intermediate steps in the workflow and visualization tools are provided to explore the results.

Analysis Workflow - Acquisition

- The first step is to read the tab-separated data file into R.
- Our raw data is a 1787-by-79 data frame. Proteins are arranged in rows and the descriptors in columns. The data-frame is containing intensity measurements, which reflect protein/peptide abundances.
- We have a total of 40 samples (4 replicates X 5 time-points X 2 conditions).

Analysis Workflow - Filtering

- Although mass spectrometry-based proteomics has the advantage of detecting thousands of proteins from a single experiment, it faces certain challenges.

- One problem is the presence of missing values in proteomics data. Look at the data and find missing values.

- One of many filtering schemes is to keep proteins that are quantified in at least two replicates in one condition.

Analysis Workflow - Normalization

- It is recommended to work with normalized data instead of intensities.
- To transform the data we can perform a global Variance Stabilization Normalization (VSN) procedure.
- VSN normalization performs a transformation similar to the log transformation.
- VSN also keeps the variance approximately constant across the whole intensity range and bringing the samples onto a same scale.

Analysis Workflow - Imputation

- There are still missing values in the filtered data.

- The statistical approach designed to deal with missing values is called imputation.

- Since missing values are associated with proteins with low levels of expression, we can substitute the missing values with numbers that are considered “small” in each sample.
  - We can define this statistically by drawing from a normal distribution with a mean that is down-shifted from the sample mean (lowest 1-th percentile, q=0.01) and a standard deviation that is a fraction of the standard deviation of the sample distribution (the median of the peptide/protein-wise standard deviations).

Analysis Workflow - Interpretation

- We can increase the interpretability of our high dimensional data by reducing the dimensionality of such datasets.

- We perform PCA analysis to interpret our available data.
  - Labelling based on experimental condition.
  - Labelling based on replicate ID.

- What do we notice?

Analysis Workflow – Batch Effects

- In proteomics, Batch Effects are technical sources of variation that confounds proper analysis inducing spurious differences between study groups.

- Batch Effects can lead to miss-interpretations of the data and as such it is important to diagnose and correct them.

- What can be potential sources for batch effects?

- `removeBatchEffect` function removes the shifts in the group means associated with the grouping factor provided (replicates), per row of the matrix.

Analysis Workflow – Differential Analysis

- Now we are ready to compare protein/peptide expression/abundance between the drug-resistant and the control lines.

- This involves performing statistical analysis on our data to extract proteins that are differentially expressed/abundant.

- Differential Analysis can be performed by comparing EGFvsPBS conditions at each time-point separately (here at time-point 60).

- Volcano plots help to visualize the differentially expressed proteins.

Integrative Analysis of Proteomics and Transcriptomics Data

Multi-Omics Rationale and Gene Expression Data

- Integrated multi-omics datasets can provide more extensive molecular insights compared individual omics.

- We can integrate the available proteomic data with transcriptomics gene expression data available online and which have been obtained on the same context (GSE6786).

- For the multi-omics analysis the previous results from the DGE and Proteomic analysis (time-point 60min) have been used and prepared (sample-wise expression + differential expression/abundances).

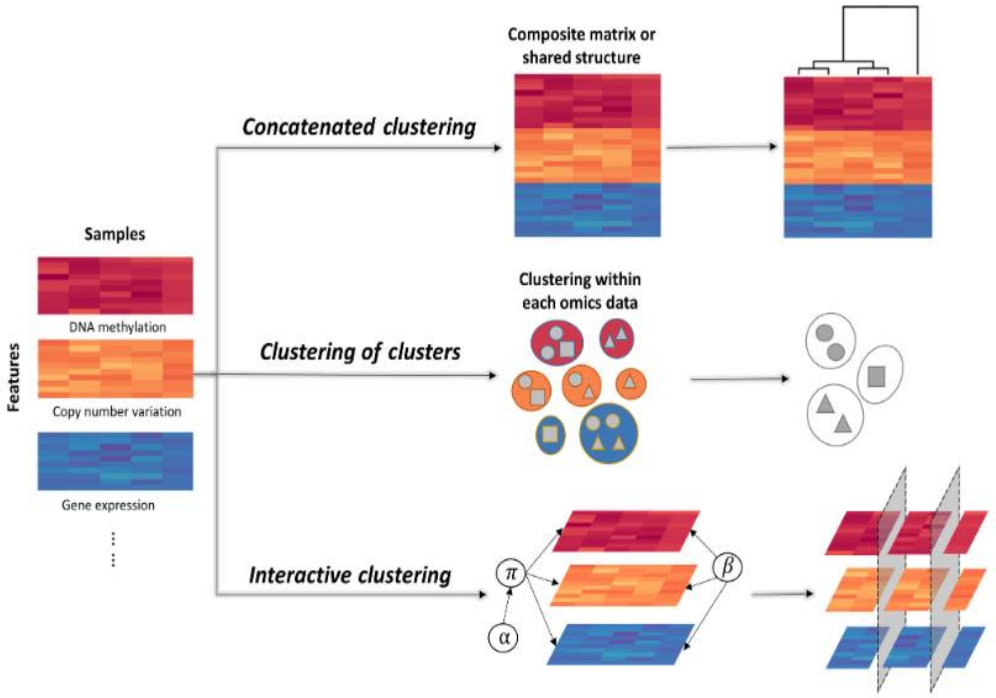
Comparison of mRNA and Protein Levels

- Relationship between mRNA and Protein abundances is not well characterised.

- We can rely on correlation analysis to measure the strength of the linear relationship between mRNA and Protein levels and compute their association.
  - Correlation at the Differential Expression/Abundance levels.
  - Correlation at the Sample-specific Expression/Abundance levels.

- What do we notice?

Clustering Analysis



- Clustering is a learning technique in which the data set is partitioned into several groups called as clusters based on their similarity or shared characteristics.
  - Typically used for the subtyping of cell-types, diseases, cohorts, etc..
- Multi-omics data can improve the accuracy and sensitivity of clustering compared individual omic layers.
- Several types of Clustering:
  - Concatenated Clustering - combine the multi-omics data into one matrix or search for the shared structure, followed by the final clustering.
  - Clustering of Clusters - Obtain the clustering information from each omics dataset first and follow by the final clustering.
  - Integrative Clustering - simultaneously integrate multi-omics data and perform clustering.

Pathway Analysis

- Gene Set Enrichment Analysis (GSEA) of pathway sets can be used to identify significantly regulated pathways based on:
  - Differential Expression values of Genes (Transcriptomic)
  - Differential Abundance values of Proteins (Proteomics)

- Genes are not the real proxies of signalling pathways and increased Protein abundance levels does not mean protein activities.

- Combined GSEA analysis at the transcriptome and proteome level can increase the accuracy of correctly identifying significantly regulated pathway sets by looking into the consensus between the two analyses.

Network Analysis

- We can apply multi-omics integrative approaches to network analysis with BioNET.

- BioNET is an R-package used to for the analysis of biological networks to identify functional modules.

- The BioNet package provides an extensive framework for integrated network analysis in R. This includes the statistics for the integration of transcriptomic and functional data with biological networks, the scoring of nodes as well as methods for network search and visualization.

Network Analysis – Getting an Interactome

- Integrated analysis of microarray/mass-spect data in the context of biological networks such as protein–protein interaction (PPI) networks has become a major technique in systems biology.

- BioNET aims at identifying functional modules (significantly differentially expressed subnetworks) within large networks.

- Where can we find large-scale resources of PPI's?
  - One option is through OmniPath.
  - Get the PPI resource with OmniPathR.

Network Analysis – Score Assignment

- We can assign a score for each node in the network reflecting its functional relevance.

- As a first step, multiple *P*-values derived from the analysis of different experiments (e.g. proteomics/transcriptomics) can be aggregated using a uniform order statistics (Fisher's method).

- Based on these aggregated *P*-values we can derive a scoring function.
  - The goal is to develop an additive score, where positive values signify signal content and negative values denote background noise.
  - A Beta Uniform Mixture (BUM) model is fitted on the entire set of aggregated p-values of all nodes in the interaction network.
  - A threshold p-value controls the FDR for the positively scoring *P*-values.

Network Analysis – Network Reconstruction

- Identification of the functional modules can be casted as finding a mathematical optimization algorithm: finding the optimal-scoring subgraph in the vertex-weighted graph.

- Maximum-Weight Connected Subgraph Problem or *MWCS*.

- Two approaches to solve such an optimization problem:
  - Stochastic approach with the *fastHeinz* algorithm.
  - Deterministic approach by formulating *MWCS* as an ILP problem.

- Try running the Stochastic approach to obtain and visualize the optimal subnetworks.