## Topic 5 - Assignment

Presentation Date: **11/11/21**

Report Submission: **11/14/21**

**Notes:**

You are free to use any programming languages/packages to perform your analysis.

For the practical session, Python 3.6 will be used along with the following package:

* cptac (pip)
* gseapy (pip)
* pandas (pip/conda)
* numpy (pip/conda)
* seaborn (pip/conda)
* matplotlib (pip/conda, version 3.1.0)

The gene set used in this assignment was generated using Independent Component Analysis (ICA). If you are interested in how the set was generated this paper describes the technique:

*Liu, Wenke, Samuel H. Payne, Sisi Ma, and David Fenyö. 2019. “Extracting Pathway-Level Signatures from Proteogenomic Data in Breast Cancer Using Independent Component Analysis.” Molecular & Cellular Proteomics: MCP 18 (8 suppl 1): S169–82.* [*https://www.ncbi.nlm.nih.gov/pubmed/31213479*](https://www.ncbi.nlm.nih.gov/pubmed/31213479)

### Tasks:

**Task 1:**

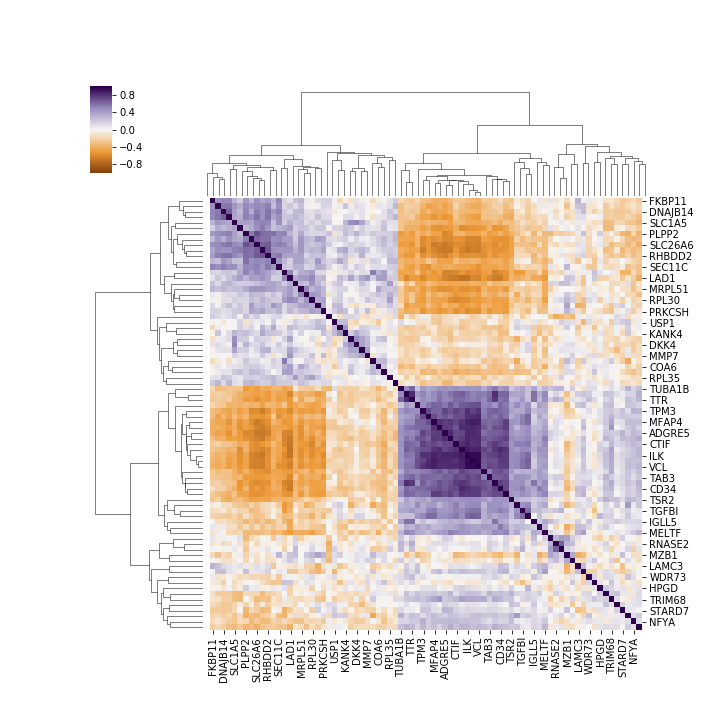
1. Using the CPTAC python package we used in class, load the **endometrial** **proteomic** data
2. Load the gene set from ‘**filtered\_genes\_list.txt’**
3. Subset the endometrial proteomics dataset to only include genes from the filtered gene list in **step 2**
4. Plot the correlation between pairs of genes in the filtered proteomics table using a heatmap (without clustering)
   1. Do no try correlating all the proteins in the original proteomics table, it is way too big
5. Use a clustermap instead of a heatmap to visualize the correlations (**Figure 1 below**)
   1. What is the significance of the resulting clustermap?

**Task 2:**

1. Using the IC\_01.rnk file, which was generated from running ICA on proteomics data on the endometrial dataset, run GSEAPreranked using the “Hallmarks” gene sets
2. Interpret the results using any visualization you chose.
   1. Are there any interesting, biologically relevant results?

INFO: ‘**IC\_01.rnk**’ was filtered to produce ‘**filtered\_genes\_list.txt**’ used in **Task1**

**Figure 1**

****