Description:

In this folder, you will find a tsv file ("Protein_abundance.tsv") that contains protein abundance measured by label-free mass-spectrometry for chronic lymphocytic leukemia (CLL) patient samples. You will also find an excel table ("sampleAnnotation.xls") that contains some basic annotations for those samples. Your task is to process the protein abundance dataset; assess its quality and identify protein markers for prognosis. You also need to write a report, preferably in **Rmarkdown or Jupyter Notebook** format, to present your analysis results to your potential future dry-lab or wet-lab collaborators.

Specific tasks:

1) **Data processing:** The protein abundance in the tsv file is not normalized and has missing values, which is very common in the data table you will get from a proteomic facility. You need to use a proper way to normalize the protein abundance and deal with missing values.

2) **Quality assessment:** The protein abundance measurement can often be influenced by technical factors, such as batch effect, operators, total protein concentrations, and free thaw cycles of the cells. Those technical factors could potentially act as confounders for downstream analysis. In the sample annotation table, you will find the technical factors and you need to evaluate whether they will confound downstream analysis.

3) **Identify protein markers for prognosis:** in the sample annotation file, you will find three columns that contain the clinical information, which can be used to estimate the overall survival, of the CLL patients. You need to select proteins whose expression can be used to predict the overall survival of those patients using a proper statistical model. You may also do an enrichment analysis to see which pathway is potentially related to clinical outcome.