**Assignment**

*Instructions:*

* *You will have 24 hours from the time of receipt to submit the solution. However, the assignment should take <= 4 hours to complete.*
* *The assignment can be completed in R or Python.*
* *For each task, we have suggested tools that you can use.*

## Background:

[RNA-Sequencing](https://en.wikipedia.org/wiki/RNA-seq) is a high-throughput method for gene expression profiling - measuring the expression of genes in a sample and revealing the presence and quantity of RNA in a biological sample. In the context of cancer, gene expression profiling can also be used to more accurately classify tumors and understand the heterogeneity within a cancer type.

In this assignment, we will work with a gene expression dataset of Pancreatic Adenocarcinoma. Pancreatic Adenocarcinoma (PAAD) is the third most common cause of death from cancer, with an overall 5-year survival rate of less than 5%, and is predicted to become the second leading cause of cancer mortality in the United States by 2030.

## Data Provided:

1. We have adataset *(file: PAAD.gct)* containing the gene expression data for pancreatic cancer samples. The data consists of expression of ~20,000 genes for 185 samples. The file format is [GCT](http://software.broadinstitute.org/cancer/software/genepattern/file-formats-guide#GCT), a tab-delimited file used for sharing gene expression data and metadata for samples.

Suggested tools to read the GCT file:

* Use R package [cmapR](https://github.com/cmap/cmapR) for reading GCTs in R
* Use python package [cmapPy](https://github.com/cmap/cmapPy) for reading GCTs in python
* Use [Phantasus](https://artyomovlab.wustl.edu/phantasus/) is an open source tool which is used to visualize GCT files, make various plots, apply algorithms like clustering and PCA among others

1. List of 25 genes defining the IFN signature (*file: type1\_IFN.txt)*
2. For Bonus question: Gene expression (log normalized TPM RNAseq counts) in a Normal Pancreas tissue *(file: Pancreas\_log\_tpm\_RNAseq\_mat.csv)*

## Tasks:

**1. Data cleaning and check the distribution of gene expression across samples**

**Steps:**

1. Remove rows (genes) with NaNs
2. Generate boxplot for gene expression for all samples.

**Questions:**

1. How many genes/rows had NaNs?
2. How is the distribution of gene expression across samples.?

**2. Subset the data for only the** [**Exocrine**](https://en.wikipedia.org/wiki/Pancreatic_cancer#Exocrine_cancers) **(adenocarcinoma) tumors and remove** [**Neuroendocrine**](https://en.wikipedia.org/wiki/Pancreatic_cancer#Neuroendocrine) **tumors based on the PCA for all samples.**

Genome and gene expression-based subtypes have been widely accepted as methods of disease stratification. We want to stratify these tumor samples by the type of pancreatic cancer they exhibit. For this, apply dimensionality reduction techniques (PCA) to find these two groups within this multi-dimensional data.

**Steps**:

1. Run PCA on the dataset and plot different samples with PCA1 and PCA2 on x-axis and y-axis. *(Suggested tools: Use python or R to generate PCA plots. You can also use a tool like* [*Phantasus*](https://bioconductor.org/packages/release/bioc/vignettes/phantasus/inst/doc/phantasus-tutorial.html)*. If you are using Phantasus, make sure to save the analysis in a JSON file and include the resulting plots in the jupyter notebook / R markdown)*
2. Overlay the information from metadata column *‘histological\_type\_other’* on top of PCA plot and check if neuroendocrine tumors are separating out.
3. Remove the Neuroendocrine tumor samples from the dataset so that it contains only the Exocrine (Adenocarcinoma) tumor samples. The histology for the different tumor samples is contained in the GCT file.

**Questions**

1. What does the analysis say about the general behavior of the different samples?
2. Are the neuroendocrine tumors clearly separable from the adenocarcinoma tumors?
3. What can be said about the variance of the PCA?

**3. Understand the effect of Interferons in Pancreatic Adenocarcinoma**

[Interferons](https://en.wikipedia.org/wiki/Interferon) (IFNs) are a group of signaling proteins made and released by host cells in response to the presence of several pathogens, such as viruses, bacteria, parasites, and also tumor cells. Type I interferons (IFNs) are a large subgroup of interferon proteins that help regulate the activity of the immune system. The genes responsible for type 1 Interferons is called [Type 1 IFN signature](https://en.wikipedia.org/wiki/Interferon_type_I) and consists a set of 25 genes in humans.

* Can you characterize the presence of IFN signature in pancreatic adenocarcinoma tumors by assigning a score to each sample which denotes the positive or negative presence of IFN genes in the sample?
* How is the distribution of this score among the different samples?
* Based on this distribution can we identify the presence of high and low IFN subtypes in the sample cohort?

**Steps:**

1. Run the GSVA (a single sample gene set enrichment) algorithm with 25 gene IFN signature as the gene set and the subsetted pancreatic cancer data as the expression dataset. (Suggested tools: Use [GSVA package](http://bioconductor.org/packages/release/bioc/html/GSVA.html)) Additional links: [GSVA paper](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-7), [A paper for reference which studies T-cell signature in PAAD](http://clincancerres.aacrjournals.org/content/23/12/3129))
2. Check distribution of GSVA scores for samples. Do the GSVA scores segregate samples into subtypes?

Installation hacks:

* Running GSVA in R: In case there are dependencies issues, install gsva inside rstudio docker container: <https://hub.docker.com/r/rocker/rstudio/>
* Running GSVA in python: Run GSVA through the docker given for gsva python <https://github.com/jason-weirather/GSVA>, <https://hub.docker.com/r/vacation/gsva>

**4. Tracking and Sharing the analysis**

A very important part of the analysis is its reproducibility. To make analysis reproducible we need to preserve threads of thoughts and share it. One way to do this is to use a framework which preserves all the steps in analysis and what else can be better than git. Fortunately, we already have solutions such as knowledge repo available. While we are doing the analysis, we would also like to share it with our colleagues and ask for their opinion.

1. Use knowledge repo to track the analysis done above and host it on GitHub (don’t make the data public). Include the background, code, plots, analysis, and your thoughts and comments in the knowledge repo.
2. ***(optional)*** Host the knowledge repo on a server (e.g. Heroku)

## Bonus Question

**5.** **Presence of IFN in Normal pancreas**

Another question which arises is about the presence of IFN signature in Normal tissue of Pancreas. The dataset provided is normalized log normalized TPM (Transcripts per million) RNA-Seq counts in a Normal Pancreas tissue.

* Check for the presence of IFN in normal tissue samples of Pancreas?
* Is GSVA a good metric to compare to compare the presence of IFN in pancreatic cancer tumors and pancreas normal tissues?
* What does the result say about the general biology involved? Is it consistent with the behavior you would expect to see with respect to immune activation in normal tissues and tumors?