Assignment

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.

Ribonucleic acid (RNA) is a polymeric molecule essential in various biological roles in coding, decoding, regulation and expression of genes. RNA and DNA are nucleic acids, and, along with lipids, proteins and carbohydrates, constitute the four major macromolecules essential for all known forms of life.

RNA-Seq (RNA sequencing), is a sequencing technique to detect the quantity of RNA in a biological sample at a given moment. Here we have a **dataset of normalized RNA Sequencing reads for pancreatic cancer tumors**. The measurement consists of ~20,000 genes for 185 pancreatic cancer tumors. The file format is GCT , a tab-delimited file used for sharing gene expression data and metadata (details for each sample) for samples.

- The R package cmapR can be used for reading GCTs in R.
- The python package cmapPy can be used for reading GCTs in python.
- Phantasus is an open source tool which is used to visualise GCT files, make various plots, apply algorithms like clustering and PCA among others.

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

- Remove genes with NaNs. How many genes had NaNs?
- Generate gene expression distribution for all samples. How is the distribution of gene expression across samples?

2. Identify only the Exocrine (adenocarcinoma) tumors and remove Neuroendocrine tumors.

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.

- Visualize the data whole data using PCA. (*Use python or R to generate the PCA plots. You may use Phantasus for validation*)
- What does the analysis say about the general behaviour of the different samples?
- Are the neuroendocrine tumors clearly separable from the adenocarcinoma tumors?-- Overlay the information from metadata column *'histological_type_other'* on top of PCA plot and check if neuroendocrine tumors are separating out.

- What can be said about the variance of the PCA?
- Which features contribute the most in PC1, PC2 and so on? Can you come up with an appropriate visualization to demonstrate the feature relevance for these PC components?
- Remove the neuroendocrine tumors from the dataset so that it contains only the adenocarcinoma tumor samples. The histology for the different tumor samples is contained in the GCT file.

Here are pointers on this exercise: pcaExplorer , Phantasus tutorial

3. Understand the effect of Interferons in Pancreatic Adenocarcinoma

Interferons (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 and consists of a set of 25 genes in homo sapiens.

- Can you plot the gene expression values for these genes for pancreatic adenocarcinoma?
- 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) Additional links: GSVA paper, A paper for reference which studies T-cell signature in PAAD)
- 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. Unsupervised analysis

Do unsupervised clustering on the gene expression data with the algorithm of your choice and identify any correlations of the clusters so formed with the sample metadata. What statistical tests do you recommend to test the strength of association between the identified clusters and the various sample metadata? Come up with appropriate visualizations.

Note: *Don't ignore the theoretical questions as they have weightage as well. Try to answer them in the notebook itself.*

Submission

Share the entire analysis which includes all plots, code and conclusions as a jupyter notebook in a private github repository. Please adhere to this form of submission and do not submit any scripts, images or word documents. Explore as much as you can and do not refrain from writing long explanations.

LINK to data folder