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 reveal the presence and 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 for samples.

- The R package cmapR can be used for reading GCTs in R.
- 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. 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.
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.

- What does the analysis say about the general behaviour of the different samples?
- Are the neuroendocrine tumors clearly separable from the adenocarcinoma tumors?
- What can be said about the variance of the PCA?

Hints: pcaExplorer , Phantasus tutorial

2. 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 a set of 25 genes in homo sapiens.

- 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 PAAD?

Hints: GSVA paper, A paper for reference which studies T-cell signature in PAAD, GSVA package

Submission

Share the entire analysis which includes all plots, code and conclusions as a jupyter notebook in a 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