**INTRODUCTION:**

I am interested to work on Pancreatic cancer and so looked for all the datasets related to it. I found the paper on Proteogenomic Characterization of Pancreatic Ductal Adenocarcinoma interesting and investigated the dataset. The paper explains about how deadliest the disease Pancreatic cancer is as the survival rate is very low. Due to the lack of early signs and symptoms along with less effective methods of tumor specific treatment, the overall survival rate achieved in patients with metastatic disease is less than 12 months.

**Project dataset:**

I have obtained my dataset from cbioportal (<https://www.cbioportal.org/datasets>). It has many files that contains clinical data, miRNA dataset, expression profiles etc.

Metadata:

The dataset involves 140 participants having 74 males and 66 females between the age of 31- 85 from 7 different countries. There are about 28,057 genes were identified in the methylation analysis. Several risk factors such as smoking history, chronic pancreatitis, obesity, type2 diabetes are associated with the disease. The dataset has clinical components such as age, sex, race, tumor site, tumor stage and vital status.

Chart

Description automatically generated

**METHODS:**

**FILE CLEANUP USING PYTHON:**

Initially, the data has been extracted from the cbiportal.org. The files for the pancreatic cancer contained clinical data and gene expression data for all the samples. The files were downloaded and cleaned by removing redundant information, junk values and altered to a format by which I could do the analysis in R. I used python for parsing and cleaning the data along with pandas for creating the data frames and SQLite database is used to generate tables. The primary files contained 28,057 genes for 140 patients. The patients were filtered to 90 by removing the patients where the live status, drinking status and smoking status is unknown. The rows were filtered by removing rows containing repeated values across all patients, also where NA’s were present. Based on the filtered patient id’s obtained from the expression file, the patients in the clinical file are filtered. In the patient files where the drinking status mentioned as past drinking status were removed. This is done because, I believed that current habits would impact the expression levels. Similarly, done for smoking status as well. The patients were ordered based on the age. Thus the two files patients file has all the filtered clinical information and the gene file has all the filtered genes.

**R ANALYSIS:**

The file 1 has the patient’s clinical data with Patients Id’s, age, smoking status, Drinking status, live status, and gender. The file 2 has the expression levels for 21,541 genes with their gene symbols names against individual patients in the columns and gene\_id’s. I chose to do linear regression as there is age variable. I did T.test for gender, smoking status and drinking status as there are two categories or groups.

**RESULTS:**

**1.Linear regression for Age :**

Null hypothesis: There is no association between gene expression and the age groups.

Alternative hypothesis: There is association between gene expression and the age groups.

When the Linear regression is performed across all the genes against the ages of the patients, the p-values generated were not uniformly distributed. Hence the null hypothesis is not true. At a false discovery rate of 10% there were no significant number of genes found.

Chart, histogram

Description automatically generated

2. **T.test for Gender:**

Null hypothesis: the difference in group means is zero.

Alternative hypothesis: There is difference in group means.

When the T.test is performed across all genes against the gender of the patient, the p-values generated is approximately uniformly distributed with a slight skew in the distribution. Hence, there is not enough evidence to reject the null hypothesis. At a False discovery rate of 10% there were 58 genes were found.

Chart, histogram

Description automatically generated

3. **T.test for Drinking Status:**

Null hypothesis: the difference in group means is zero.

Alternative hypothesis: There is difference in group means.

When the T.test is performed for all the genes versus Drinking status, the p-values generated were not uniformly distributed. Therefore, we reject the null hypothesis. At a False discovery rate of 10% there are no genes found.

Chart, histogram

Description automatically generated

4. **T.test for Smoking Status:**

Null hypothesis: the difference in group means is zero.

Alternative hypothesis: There is difference in group means.

When the T.test is performed for all the genes with smoking status, the p-values generated were not uniformly distributed. Therefore, we reject the null hypothesis. At a False discovery rate of 10% there are 8 genes found.

Chart, histogram

Description automatically generated

**CONCLUSION:**

By doing these analysis, I could find that from the pancreatic cancer dataset, the p values generated by linear Regression for Age, T.test for drinking and smoking test displayed not uniform distribution. The t.test for gender displayed somewhat uniform distribution.

**References:**

1. Dataset portal - [**https://www.cbioportal.org/datasets**](https://www.cbioportal.org/datasets)

2. Dataset- <https://www.cbioportal.org/study?id=paad_cptac_2021>

3. Paper related to the dataset- <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8654574/>