CAPSTONE PROJECT

TITLE: Identifying drug targets for pancreatic ductal adenocarcinoma (PDAC) using Gene Expression Data

Submitted by

A. Cathy Jemimah

1. INTRODUCTION

1.1. BACKGROUND INFORMATION:

Pancreatic ductal adenocarcinoma (PDC) is one of the most lethal cancers worldwide and the aggressive form of pancreatic cancer, accounting for nearly 90% of all cases. It arises from the ductal cells of the pancreas, which are responsible for producing the enzymes that help digest food. Risk factors include having a family history of the disease, history of chronic inflammation of the pancreas (pancreatitis), Lynch syndrome, diabetes, being overweight or obese, and smoking. Unfortunately, PDAC is often diagnosed at an advanced stage due to its lack of early symptoms and aggressive nature. Once diagnosed, the prognosis is often poor, with a five-year survival rate of only about 8%. Current treatment options, such as chemotherapy and radiation, have limited success in improving overall survival rates. This cancer also exhibits genetic complexity, with various mutations and alterations. Thus, there is a critical need for novel and more effective therapeutic agents. The pancreatic tumour microenvironment plays a crucial role in cancer progression and resistance to treatment. So drug development should focus on understanding and targeting the components of the tumour microenvironment to enhance the effectiveness of therapeutic interventions.

Addressing a disease like PDAC with knowledge from the gene level using expression data serves as a potential tool for treatment of this disease. Analyzing gene expression data provides a deeper insight into genes that serve as potential drug targets. It can aid in the discovery of potential biomarkers that serve as indicators of disease progression, prognosis, or treatment response. Identifying reliable biomarkers and the corresponding genes are essential for developing targeted therapies and predicting patient outcomes.

1.2. OBJECTIVES:

- Obtain gene expression data:

  To know more about the gene expression, the samples between tumor and non-tumor tissue samples can be compared. This significantly helps in identifying the genes that are significantly up-regulated or down-regulated in tumors.

- Perform pathway enrichment:

  To perform pathway enrichment analysis for the identified differentially expressed genes using DAVID. This is done to visualize enriched terms and their relationships among the genes.
- Identify dysregulated pathways:

  To analyze the functions and pathways associated with the identified DEGs using gene ontology databases and pathway analysis tools like STRING, Cytoscape etc. Important and highly differentiating genes are identified as target.
- Identify potential drug targets:

  Focus on DEGs encoding proteins with potential characteristics such as known binding sites, structural feasibility, and novelty. Utilize existing drug-target databases and network analysis tools to identify known or predicted interactions between DEGs and existing or potential drugs.

2. METHODOLOGY

2.1. DISEASE AND SCOPE

2.1.1. Disease Selection:

Pancreatic adenocarcinoma prevails to be the 4th leading cause for cancer worldwide, with an average death count of 4, 96,000 annually along with a survival rate of around 8.5%. Being on the complex tumours, it has significant intratumoral heterogeneity. Existing therapies like surgery, chemotherapy, and radiation often show limited efficacy due to drug resistance and disease recurrence. Keeping in mind the scope available for alternative therapies, this cancer was chosen for this project. By understanding the functional roles of DEGs, we can prioritize potential drug targets and develop more effective therapies. Identifying gene expression signatures could aid in early diagnosis, prognosis prediction, and personalized treatment strategies. The high prevalence, poor prognosis, and complex nature of pancreatic adenocarcinoma combined with the availability of rich gene expression data make it an ideal candidate for analysis. Studying its molecular landscape holds immense potential for uncovering novel drug targets, developing effective therapies, and ultimately improving patient outcomes.

2.1.2. Research Question:

Can we identify novel and druggable targets for pancreatic adenocarcinoma by integrating differentially expressed genes analysis with functional analysis and network-based approaches?

2.2. DATA ACQUISITION AND PROCESSING

2.2.1. Data Sources:

Various databases are available for obtaining gene expression data for cancer including NCBI GEO (Gene Expression Omnibus) and TCGA (The Cancer Genome Atlas). Here for this study, I have taken my gene expression dataset from NCBI GEO website with the accession ID: GSE28735. This data type was a Microarray gene expression data titled as "Microarray gene-expression profiles of 45 matching pairs of pancreatic tumor and adjacent non-tumor tissues from 45 patients with pancreatic ductal adenocarcinoma" .The sample size was 90 which consist of 45 matching pairs of tumor and adjacent non-tumor tissues taken from PDAC patients. The patient population consists of both male and female with an age group ranging from 40 to 80.

2.2.2. Data Cleaning and Pre-processsing:

The data provided by GEO was a raw data which was in zipped format with all the sample information. Raw data is not suitable for further analysis. So data must be cleaned and pre-processed before analysis. This helps in reducing the noise in the dataset. For further analysis, R language is used as it is much easier when working with larger dataset with lots of information. Packages like GEOquery play a crucial role in accessing and analyzing gene expression data from the Gene Expression Omnibus (GEO); dplyr, tidyverse are other some packages utilized in this process.

2.2.3. Data Exploration:

Exploratory data analysis (EDA) plays a crucial role in understanding the distribution of the data and thus helping in quality control of the dataset. It also helps in exploring the relationships between the samples in the dataset. For EDA, various plots like box and whisker plot, density plot etc.

❖ **BOX AND WHISKER PLOT**:

This plot helps in identifying if there are any outliers in the dataset. Box plot talks about the distribution of the dataset based on the median values. Median values are utilized in this plot.

INTERPRETATION: The box plot compares gene expression in tumor samples (green box) and non-tumor samples (purple box).This plot contains y-axis which shows the normalized gene expression values and x – axis shows the sample information. Outliers are observed as circles found beyond the whiskers, which in this case is null. So there are no outliers in this plot. The distribution of expression values is skewed to the right in both tumor and non-tumor samples, meaning there are more data points towards the higher end of the range. The horizontal line within each box represents the median expression value. The median expression value is higher in tumor samples than in non-tumor samples. The box represents the interquartile range (IQR), which is the range that contains the middle 50% of the data points. The IQR is larger in tumor samples than in non-tumor samples, indicating that there is more variability in gene expression in tumor samples.

✓ Commands used:

- ex <- exprs(gset)

# box-and-whisker plot

- dev.new(width=3+ncol(gset)/6, height=5)

#Order samples by group

- ord <- order(gs)
- palette(c("#1B9E77", "#7570B3", "#E7298A", "#E6AB02", "#D95F02", "#66A61E", "#A6761D", "#B32424", "#B324B3", "#666666"))
- Par (mar=c(7,4,2,1))
- title <- paste ("GSE28735", "/", annotation(gset), sep ="")
- boxplot(ex[,ord], boxwex=0.6, notch=T, main=title, outline=FALSE, las=2, col=gs[ord])
- legend ("topleft", groups, fill=palette(), bty="n")

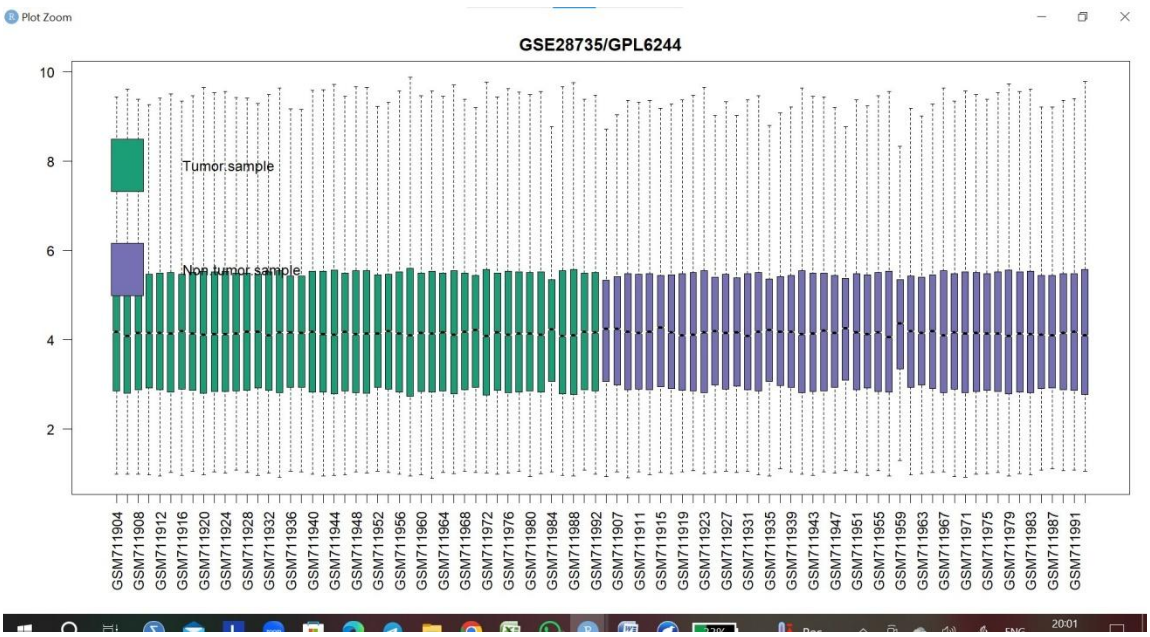

Figure 1 - BOX PLOT obtained from R analysis

- ## <u>**DENSITY PLOT**</u>:

This plot is primarily used to understand the distribution of continous data over a period of time by using kernel density estimation. This smoothens out the data, revealing trends and patterns that might be obscured by binning. Density plot helps in identifying the central tendency and skewness of the data.

<u>INTERPRETATION</u>: Here, in this plot, the absence of outliers is yet again confirmed based on the peak. In density plot, x-axis represents the gene expression values and y-axis represents the corresponding density. The distribution of expression values is skewed to the right in both tumor and non-tumor samples, meaning there are more data points towards the higher end of the range. The density of the distribution is higher in the middle of the range for both tumor and non-tumor samples, and lower at the extremes. This means that there are more samples with expression values that are close to the median, and fewer samples with expression values that are very high or very low. The tails of the distribution are longer for tumor samples than for non-tumor samples. This means that there are more tumor samples with expression values that are very high or very low, compared to non-tumor samples. This suggests that there may be more variability in gene expression in tumor samples. Based on the plot, cumulative distribution function (CDF) values can be calculated, which here tells us the median of the dataset. Shape and peaks of the graph helps us understand the distribution better. Here, in this plot, the shape is symmetrical – bell shaped distribution, which indicates normal distribution. Peak is found between 0.15 and 0.20 which is found to be the median value.

✓ Commands used:

# Expression value distribution

- par(mar=c(4,4,2,1))
- title <- paste ("GSE28735", "/", annotation(gset), " value distribution", sep ="")
- plotDensities(ex, group=gs, main=title, legend ="topright")

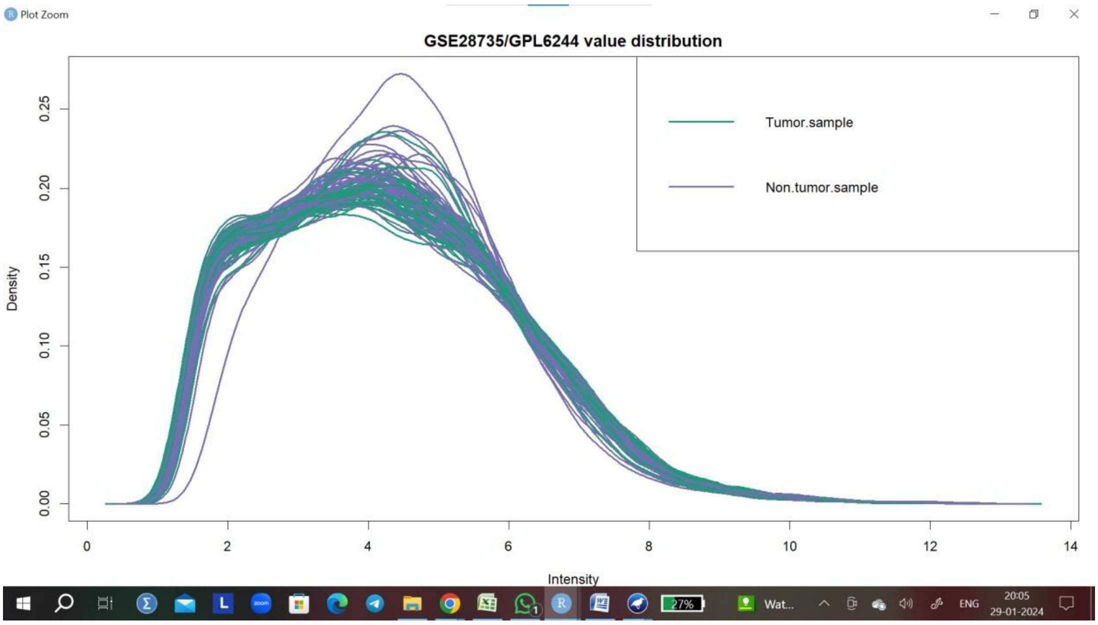

Figure 2 - DENSITY PLOT obtained from R Analysis

2.3. DIFFERENTIAL GENE EXPRESSION ANALYSIS:

2.3.1. Statistical Methods:

Different types of statistical methods are employed to understand the distribution of gene expression data. In this microarray dataset, **moderated t-test** is performed for hypothesis testing.

✓ Commands Used:

# compute statistics and table of top significant genes

- fit2 <- eBayes(fit2, 0.01)
- tT <- topTable(fit2, adjust="fdr", sort.by="B", number=250) - explain the command

HYPOTHESIS:

- **Null Hypothesis (H_0)**:

According to null hypothesis, gene expression levels are the same across groups, i.e., Tumor vs Non-tumor samples.

- **Alternate Hypothesis (H_a):**

Alternate hypothesis states that gene expression levels are different across the mentioned groups

So in order to disapprove null hypothesis based on p-value, we need to perform statistical tests. For this microarray dataset, a library called **LIMMA (Linear Models for Microarray Data**) is used. It considers experimental design and addresses variance issues, offering more accurate results.

✓ Commands used for DEG Analysis:

# Skip missing values

- gset <- gset[complete.cases(exprs(gset)),]

# Fit linear model

- fit <- lmFit(gset, design)

# Set up contrasts of interest and recalculate model coefficients

- cts <- paste(groups[1], groups[2], sep="-")
- cont.matrix <- makeContrasts(contrasts=cts, levels=design)
- fit2 <- contrasts.fit (fit, cont.matrix)

# Compute statistics and table of top significant genes

- fit2 <- eBayes(fit2, 0.01)
- tT <- topTable(fit2, adjust="fdr", sort.by="B", number=250)

So by using the above commands, genes with p values less than 0.05 were observed. From this it is clear that null hypothesis is incorrect. And from this we know that gene expression levels are different across groups, i.e., Tumor vs Non-tumor samples. This dataset consists of about 28,869 genes from which genes with p values less than 0.05 are filtered and a gene count of 10,926 obtained. Also to obtain significant genes, top 271 genes were selected based on logFC values greater >1.

2.3.2. Visualization:

In order to confirm the presence of differentially expressed genes (271 DEG), visualizing them in efficient plots are done. This gives a clear picture of how the genes in the dataset are distributed.

VOLCANO PLOTS:

Volcano plots are most commonly used to visually identify significant and meaningful changes in gene expression. It quickly focuses on genes with both strong statistical evidence and large effect sizes.

INTERPRETATION: This volcano plot only shows genes that have a statistically significant difference in expression between the two groups, based on a p-value threshold of 0.05. Each dot on the plot represents a gene. Only genes with a p-value less than 0.05 are shown, which means that there is strong evidence that their expression is different between tumor and non-tumor samples. The x-axis shows the fold change in gene expression between tumor and non-tumor samples. A positive fold change means that the gene is more highly expressed in tumor samples, while a negative fold change means that it is more highly expressed in non-tumor samples. The y-axis shows the -log10 (p-value). Higher values on the y-axis indicate stronger statistical evidence for a difference in expression between tumor and non-tumor samples. The genes that are furthest to the right (positive fold change) and highest up (most negative -log10(p-value)) are the most likely to be truly differentially expressed between tumor and non-tumor samples. These genes are the most interesting candidates for further study.

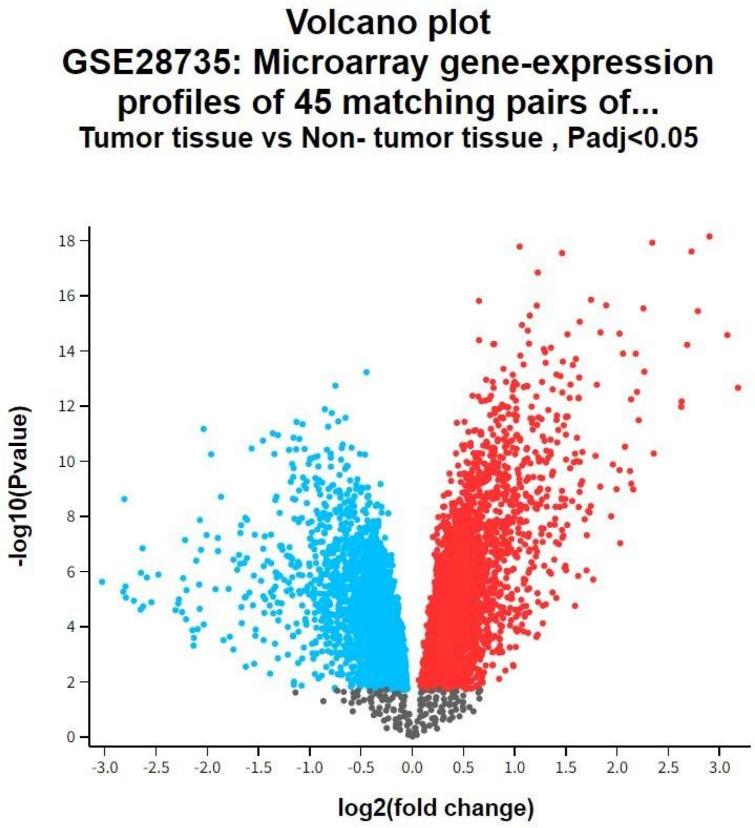

Figure 3 - VOLCANO PLOT based on p value < 0.05

✓ Commands Used:

```
# volcano plot (log P-value vs log fold change)
```

- colnames(fit2) # list contrast names
- ct <- 1   # choose contrast of interest
- volcanoplot(fit2, coef=ct, main=colnames(fit2)[ct], pch=20,
        highlight=length(which(dT[,ct]!=0)), names=rep('+', nrow(fit2)))

<u>Q-Q PLOT</u>:

A Q-Q plot is a graphical method used to assess whether the observed data follows a specific theoretical distribution, such as the normal distribution. This normal quantile-quantile (Q-Q) plot compares the quantiles of our data to the quantiles of a normal distribution. In this case, it is comparing the distribution of the moderate t-statistic obtained from a comparison of tumor and non-tumor samples to a normal distribution.

INTERPRETATION: In this plot, the points in the plot deviate from the diagonal line in a way that suggests that the distribution of the moderate t-statistic is not perfectly normal. There are more points above the line than below, which suggests that the tails of the distribution are heavier than those of a normal distribution. This means that there are more extreme values (both positive and negative) than would be expected in a normal distribution. The points deviate from the line more in the left tail than in the right tail. This suggests that there are more negative t-statistics than would be expected in a normal distribution.

✓ Commands used:

# create Q-Q plot for t-statistic

- t.good <- which(!is.na(fit2$F)) # filter out bad probes
- qqt(fit2$t[t.good], fit2$df.total[t.good], main="Moderated t statistic")

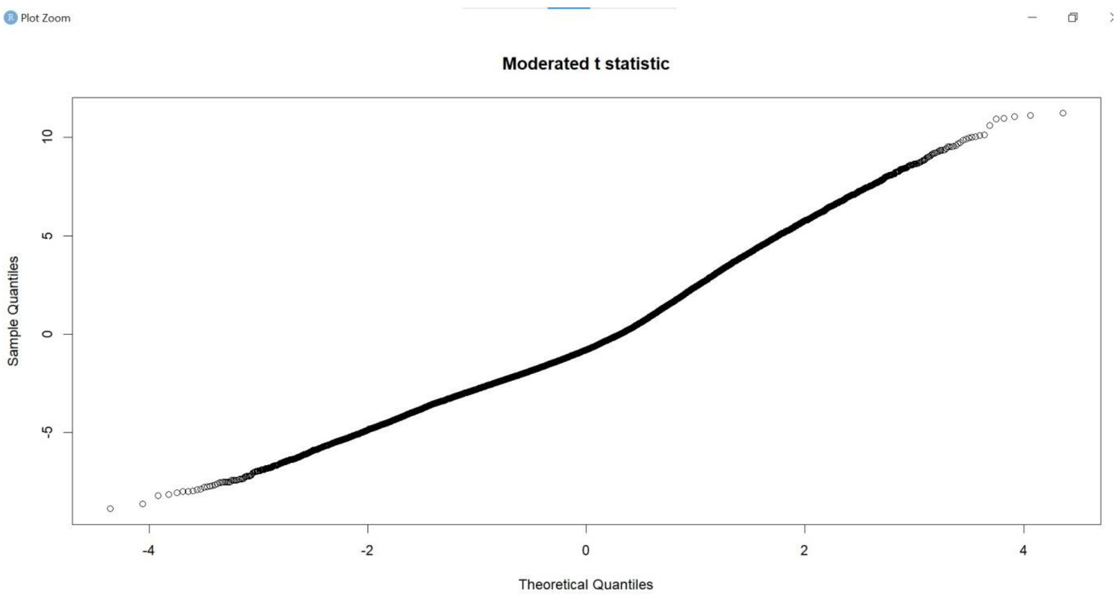

Figure 4 : Q-Q Plot obtained from R analysis

2.4. Drug target prioritization:

2.4.1. Functional Analysis:

Functional analysis of gene helps in understanding biological effects of gene expression. It can reveal unexpected connections between differentially expressed genes, suggesting novel hypotheses about how they might interact or contribute to the biological phenomenon. KEGG Pathway is a valuable resource for analyzing Differentially Expressed Genes (DEGs) and understanding the potential biological processes and pathways affected by their expression changes. It covers diverse molecular interactions, pathways, and metabolic networks across various organisms. Similar to KEGG pathway, Reactome pathway, IPA are also used for pathway and functional analysis.

DAVID SOFTWARE:

Several tools and software are utilized for functional analysis, especially DAVID (Database for Annotation, Visualization and Integrated Discovery). It is a powerful tool for analyzing the functional enrichment of gene lists, making it valuable for exploring the potential functions and pathways associated with differentially expressed genes (DEGs). This database helps in identifying enriched functional groups of genes based on GO, KEGG, and other annotation sources.

DAVID USAGE PROTOCOL:

Using the gene ID obtained from the significant gene count of **271**, DAVID was used for functional and pathway analysis. Gene ID was copied and pasted in DAVID server by using the identifier to be ‘Official gene symbol’. After analysis, it was found that there are **15 KEGG pathways** available for the given genes. From the mentioned pathways select the ones that are in close relation with the selected disease (Pancreatic adenocarcinoma). **4 pathways** were shortlisted are a total of **32 genes** were identified to be functionally important in different pathways. These genes can be utilized for further network analysis.

2.4.2. NETWORK ANALYSIS:

Network analysis has become a powerful tool in drug target prioritization of differentially expressed genes (DEGs), offering valuable insights beyond solely identifying genes with altered expression levels. Various tools like Cytoscape, string are used for network analysis.

STRING DATABASE:

The STRING database (Search Tool for the Retrieval of Interacting Genes/Proteins) is a valuable resource for network analysis. It gives us a visual representation of how genes interact with each other. From my analysis, **32** shortlisted genes were chosen for network analysis. The gene ID were written as a list and then uploaded, by giving the appropriate organism name. So from this analysis, I got a mapped network with nodes and edges mapped. From the resultant network, we can identify highly potential genes based on the interaction among them. MET gene is found to have more interactions with other genes, comparitively. After MET, genes like PLAT, LAMB3 genes were observed to have more interactions.

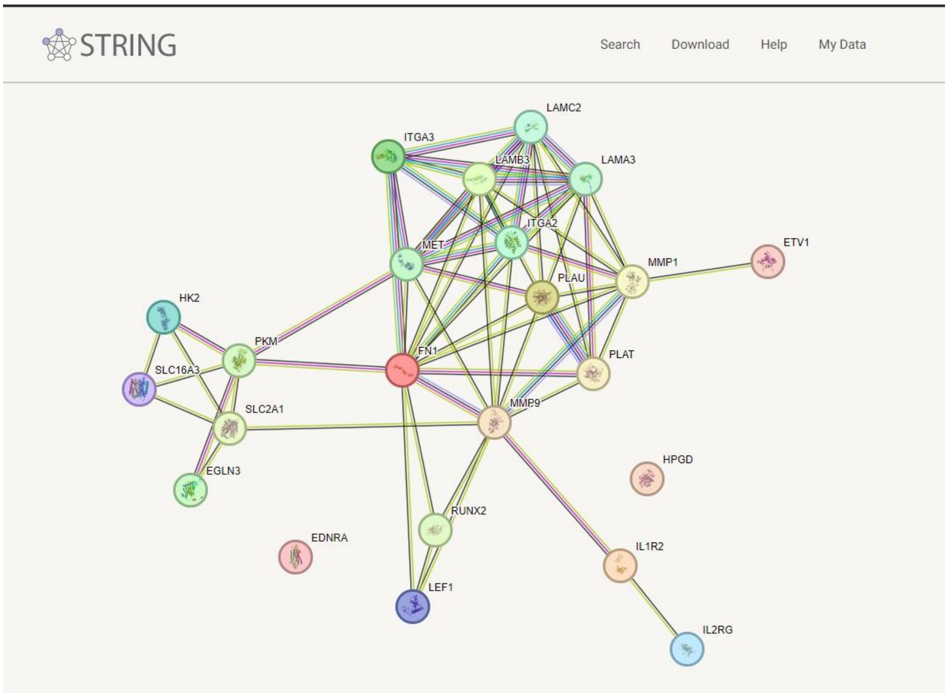

Figure 5 : Network obtained from STRING Database

DGIdb DATABASE:

To know whether the genes that we selected have been used as a potential for drug target, databases like DGIdb was used. DGIdb(Drug Gene Interaction Database) is an free, open-source database that integrates information on drug-gene interactions and druggable genes from multiple sources. Using this database, I copied the gene list and found out the druggability of the selected genes. From this it was observed that MET gene had the most druggable potential among the other genes given. Along with MET, IL2RG, LAMB3 and LAMC2 genes also had the druggability potential.

3. ACKNOWLEDGEMENT:

"Gratitude is not only the greatest of virtues, but the parent of all others." - Marcus Tullius Cicero

First and foremost, I praise the God Almighty for pouring his grace abundantly and helping me complete this course and project. I would like to extend my heartfelt thanks to the Bversity team, for curating such an amazing course with utmost care and concern. I appreciate their invaluable support and guidance throughout this course and especially this project. Their expertise, encouragement and constant encouragement were instrumental in my success. In particular, I would like to thank Mr. Sudarsan Varadharajan, Aravind N and their entire team for arranging such engaging classes. Special mention to Ms. Sakthi and their team,Your growth factor, for helping us get along with this course by serving as a bridge between mentors and students. I am also deeply grateful to my mentors, Dr. Sudeesh . K. Prabhudas, Dr. Anuranjan Singh Rathore, Dr. Zaiba Hasan Khan and Dr. Manisha Bharadwaj for their insightful feedback, technical assistance, and unwavering belief in my capabilities. Their mentorship played a crucial role in shaping my understanding of the subject matter and refining my research approach. Finally, I would like to express my gratitude to Mrs. Bhuvaneshwari, Dr. Vinoth , Mrs. Subha from Sona College of Arts and Science, for their careful review and feedback on my project. Their expert validation has significantly enhanced the quality of my work. Without the help and support of these individuals, this project would not have been possible. I am truly grateful for their contributions and am excited to continue learning and growing in the future.

4. RESULTS AND COMMUNICATION:

4.1. Findings:

From DGE analysis of PDAC dataset, by comparing pancreatic tissue and adjacent tissue samples, I found out that there were 4 genes which have the ability of acting as a drug target. The other 3 genes also were found to play a crucial role in changing the tumor microenvironment.

S.No.	GENE	ROLE IN PANCREATIC ADENO CARCINOMA	SUSCEPTIBILITY AS DRUG TARGET FOR PDAC
01.	MET	- Tumor progression and metastasis: Promotes cell proliferation, survival, invasion, and migration. - Signaling pathway: Key component of the HGF/MET signaling pathway, which is often dysregulated in pancreatic cancer.	Potentially high: - Several MET inhibitors are in clinical trials for pancreatic cancer, with some promising results
02.	IL2RG	- Immune response: Plays a role in T cell activation and immune response regulation. - Tumor microenvironment: Can be expressed by tumor cells and immune cells in the tumor microenvironment.	Moderately high: - Some ongoing research exploring targeting IL2RG in cancer, but more evidence is needed to determine its potential as a therapeutic target in pancreatic cancer.
03.	LAMB3	- Tumor invasion and metastasis: May promote tumor invasion and metastasis by regulating cell adhesion and migration	Moderately high: - LAMB3 is being investigated as a potential biomarker and therapeutic target in pancreatic cancers - Targeting the ECM is an emerging area of cancer

			research, and LAMB3 could be a potential target within this approach.
04.	LAMC2	- ECM composition and remodeling: Similar to LAMB3, contributes to the structure and function of the ECM - Tumor progression and metastasis: May play a role in tumor progression and metastasis by regulating cell adhesion and migration.	Moderately high: - Similar to LAMB3, research on LAMC2 as a therapeutic target in pancreatic cancer is limited, but it holds potential within the broader strategy of targeting the ECM.

4.2. Limitations:

Though this is a highly potential method for analysis, there are certain limitations which must be taken care of.

❖ **DEG Analysis Limitations**:

- **Tumor heterogeneity**: PDAC tumors are highly heterogeneous So, DEG analysis based on bulk tumor samples may not capture this heterogeneity and miss important differences in specific subpopulations of cells.
- **Confounding factors**: Many factors besides cancer can influence gene expression, such as inflammation, stromal cells, and patient variability. DEG analysis may identify genes that are differentially expressed due to these factors rather than being directly involved in tumor formation.
- **Limited information**: DEG analysis focuses on differences in gene expression levels, but it doesn't tell us about protein activity, post-translational modifications, or interactions with other molecules, which are essential for understanding their role in cancer.

- **Drug Target Identification Limitations**:
  - **Correlation ≠ causation**: Just because a gene is differentially expressed in cancer doesn't mean it's a good drug target. It could be a passenger mutation with no functional role in tumor formation.
  - **Drug development challenges**: Even if a promising target is identified, developing a safe and effective drug that specifically targets it can be expensive, time-consuming, and have a low success rate.
  - **Resistance**: Pancreatic cancer is notorious for developing resistance to therapies, and targeting a single gene is unlikely to be a long-term solution.
- **Need for Further Research and Validation**:
  - **Single-cell analysis**: Studying gene expression on the single-cell level can help overcome tumor heterogeneity and identify relevant subpopulations of cells.
  - **Preclinical and clinical validation**: Promising targets identified through DEG analysis need rigorous testing in preclinical models and eventually clinical trials to assess their safety and efficacy in patients.

5. CONCLUSION:

This project was focused on identifying potential drug targets for pancreatic cancer by differential gene expression analysis, which was done using R analysis and further analysed using tools like DAVID, KEGG Pathway, STRING and DGIdb. From a gene count of 28,869, after several steps like pre-processing, statistical analysis, the gene count reduced to 271. From these genes a total of 32 genes were identified in KEGG pathway. From those genes 4 genes (MET, IL2RG, LAMB3, and LAMC2) showed high potential in serving as drug target based on gene-gene interaction and network analysis. MET gene emerged as a potential drug target due to its established role in PDAC progression and ongoing research into MET inhibitors. While MET is a promising target, further research is crucial to confirm its functional role in your specific PDAC context. Considering the heterogeneity of PDAC, investigating MET expression and function at the single-cell level could reveal subpopulations of cells particularly dependent on MET signalling, offering more specific therapeutic targets. Combining DEG data with other information like protein-protein interactions or pathway analysis might uncover additional genes or pathways downstream of MET that could be targeted for synergistic effects.