CAPSTONE PROJECT

TITTLE: Identify drug targets for Diabetic Retinopathy using gene expression analysis

Submitted by

M.Chandru

1. Introduction

1.1. Background information:

Diabetic mellitus, also commonly known as diabetes, is a chronic metabolic disorder characterized by high blood sugar (glucose) levels. This occurs when the body either doesn't produce enough insulin, a hormone that regulates blood sugar, or the body's cells become resistant to its effects. One of the complications of diabetic mellitus is Diabetic retinopathy. It is a serious eye disease that is caused by high blood sugar levels damaging the blood vessels in the retina. It is the light-sensitive layer of tissue at the back of the eye.

Stages of diabetic retinopathy:

There are two main stages of diabetic retinopathy: non-proliferative diabetic retinopathy (NPDR) and proliferative diabetic retinopathy (PDR). In the early stage (NPDR), the blood vessels in the retina become weak and leaky. This can cause symptoms like blurred vision, floaters (spots in your vision), and difficulty seeing at night. In the more advanced stage (PDR), new, abnormal blood vessels grow in the retina. These new blood vessels are fragile and can bleed or leak fluid, which can lead to severe vision loss or even blindness.

Significance in Drug Discovery:

Drug discovery is a complex and lengthy process with no guaranteed outcomes. It requires Prevalence, Impact, Limited Treatment Options and Need to cure for the disease. Diabetes affects millions globally; Diabetic retinopathy is a common complication of diabetes, affecting around 30% of people with Type 1 diabetes and 60% of those with Type II diabetes within 20 years of diagnosis. DR leads to Vision loss and can have devastating consequences for quality of life, independence, and employability. There are less treatment for DR, existing therapies manage symptoms but don't prevent or cure DR.

Importance of Gene Expression Data:

Gene expression profiles can offer valuable insights into the molecular mechanisms underlying DR. By analyzing how genes are expressed or silenced in healthy and diseased retinas. Understanding the pathways involved in DR pathogenesis helps pinpoint potential drug targets. Genes with altered expression could become biomarkers for early diagnosis, disease progression monitoring, and personalized treatment. Genes encoding crucial proteins in pathological processes become potential targets for therapeutic intervention. Examples: Studies have identified genes involved in inflammation, angiogenesis, and vascular permeability – all contributors to DR. Gene expression analysis led to the development of drugs targeting VEGF, a growth factor promoting abnormal blood vessel growth in DR.

1.2. Objectives:

1. Gene expression data acquisition:

To search databases like NCBI GEO for required gene expression data of DR affected individuals and utilizing them for further analysis. This helps to understand the disease mechanism by potentially identifying key proteins and their pathways which gives a clear picture of deregulation of the disease.

2. Analyze gene expression data:

To compare, process and interpret the selected gene expression profiles between DR patients and healthy controls and to identify genes with significant differential expression which can be implemented using tools like R and python.

3. Identify deregulated pathways:

Construct protein-protein interaction (PPI) networks based on the identified DEGs and identify hub genes with high connectivity in DR pathogenesis. Mapping of such genes to relevant biological pathways using tools like Cytoscape is to be done and identify the potential target areas for drug action.

4. Priorities potential drug targets:

To identify potential drug targets based on their functional relevance, novelty and the ability for therapeutic benefit targeting the inhibition of identified differential gene expression. To understand the mechanisms of action for proposed drugs and their potential impact on DR patho-physiology is crucial.

2. Methodology

2.1 Define the Disease and Scope:

2.1.1 Disease Selection:

Diabetic retinopathy (DR) affects a significant portion of the global population. The International Diabetes Federation estimates that 93 million adults have DR worldwide, with 28.8 million experiencing vision-threatening stages. Untreated DR can cause vision loss and blindness. The prevalence of diabetes, and consequently DR, is projected to rise continuously in the upcoming decades, further amplifying the need for effective interventions.

2.1.2 Research Question:

Can differential gene expression analysis combined with network based approaches help in discovering potential binding site of specific genes to target in the prevention of diabetic retinopathy affecting the diabetes mellitus patients?

2.2 Data Acquisition and Processing:

2.2.1 Data Sources:

NCBI GEO database is used for accessing suitable dataset for this project. The dataset with the accession id: **GSE221521** was chosen for analysis. This GEO dataset delves into the influence of diabetic retinopathy (DR) on peripheral blood gene expression. This publicly available resource offers valuable insights into the systemic factors potentially linked to this sight-threatening complication of diabetes mellitus. The dataset includes normalized gene expression data, along with clinical information for each participant.

The dataset compares gene expression profiles in blood samples from three groups:

- 50 healthy controls
- 74 diabetic patients without diabetic retinopathy (DM group)
- 69 diabetic patients with diabetic retinopathy (DR Group)

2.2.2 Data Cleaning and Preprocessing:

Since my dataset was already normalized, there was no need to do data cleaning and pre-processing to find any missing values and outliers.

2.2.3 Data Exploration:

A box plot is a useful visualization for exploring the distribution of a numerical variable across different groups or categories.

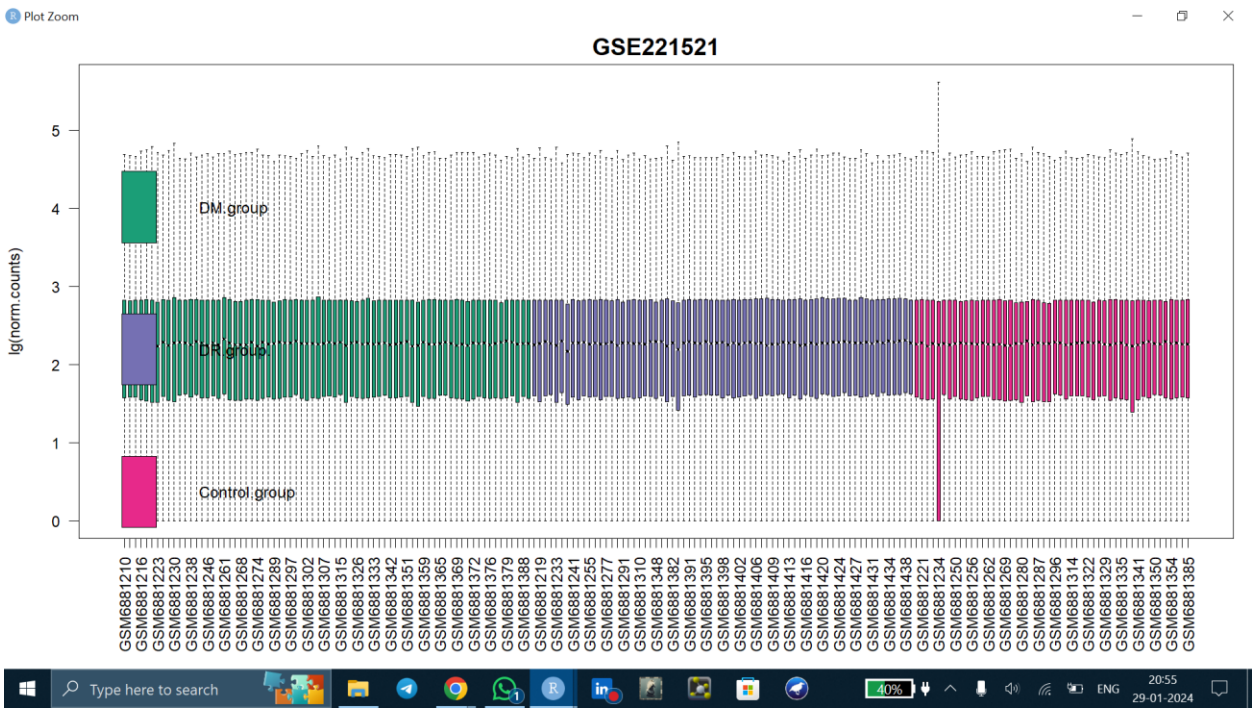

Figure 1: Box and Whisker plot

This is a boxplot generated from gene expression data from the GSE221521 dataset on NCBI GEO. The boxplot compares the expression of a specific gene across three groups:

- Healthy controls (Ctrl),
- Diabetic patients without diabetic retinopathy (DM),
- Diabetic patients with diabetic retinopathy (DR).

Distribution of expression:

- The center line of each box represents the median expression level of the gene in each group.
- The box represents the interquartile range (IQR), encompassing the middle 50% of the data points.
- The whiskers extend to the most extreme data points within 1.5 times the IQR from the median.
- There are no outliers visible in the plot.

Group comparisons:

- Visually, the boxes for the Ctrl and DM groups appear to have some overlap, suggesting some similarity in expression levels.
- The DR group's box seems to be shifted slightly to the right, indicating potentially higher expression levels compared to the other groups. However, due to the overlap, it's difficult to say definitively if this difference is statistically significant.

2.3 Differential Gene Expression Analysis:

2.3.1 Statistical Methods:

I have chosen the Likelihood Ratio Test (LRT) for your differential gene expression analysis.

LRT:

The LRT is a parametric test used to compare two nested models and assess whether the more complex model provides a significantly better fit to the data compared to a simpler model. In the context of RNA-seq data, the models represent different hypotheses about gene expression levels:

- Null hypothesis (H0): Expression levels are the same across groups.
- Alternative hypothesis (Ha): Expression levels differ between groups.

The LRT compares the likelihood of the data under each model and calculates a test statistic based on the difference in log-likelihoods. This statistic is then compared to a chi-square distribution with degrees of freedom equal to the difference in the number of parameters between the two models.

Statistical significance:

I mentioned that the p-value for this comparison is greater than 0.05. This means that we cannot reject the null hypothesis that the expression levels are similar across the groups with statistical confidence.

Commands used for DEG analysis:

#Using "DESeqDataSetFromMatrix" to form a group of count data and column data

- ds <- DESeqDataSetFromMatrix(countData=tbl, colData=sample_info, design= ~Group)

#Use LRT for all-around gene ranking

- ds <- DESeq(ds, test="LRT", reduced = ~ 1)

This command gives a result of LRT test.

2.3.2 Visualization

1. Volcano plot:

A volcano plot is a type of scatter plot commonly used in differential gene expression analysis to visualize the relationship between fold change and statistical significance for each gene.

After extracting results for top genes which has expressed we can visualize using volcano plot.

```
# volcano plot

old.pal <- palette(c("#00BFFF", "#FF3030")) # low-hi colors

par(mar=c(4,4,2,1), cex.main=1.5)

plot(r$log2FoldChange, -log10(r$padj), main=paste(groups[1], "vs", groups[2]),

xlab="log2FC", ylab="-log10(Padj)", pch=20, cex=0.5)

with(subset(r, padj<0.05 & abs(log2FoldChange) >= 0),

points(log2FoldChange, -log10(padj), pch=20, col=(sign(log2FoldChange) + 3)/2,
cex=1))

legend("bottomleft", title=paste("Padj<", 0.05, sep=""), legend=c("down", "up"),
pch=20,col=1:2)
```

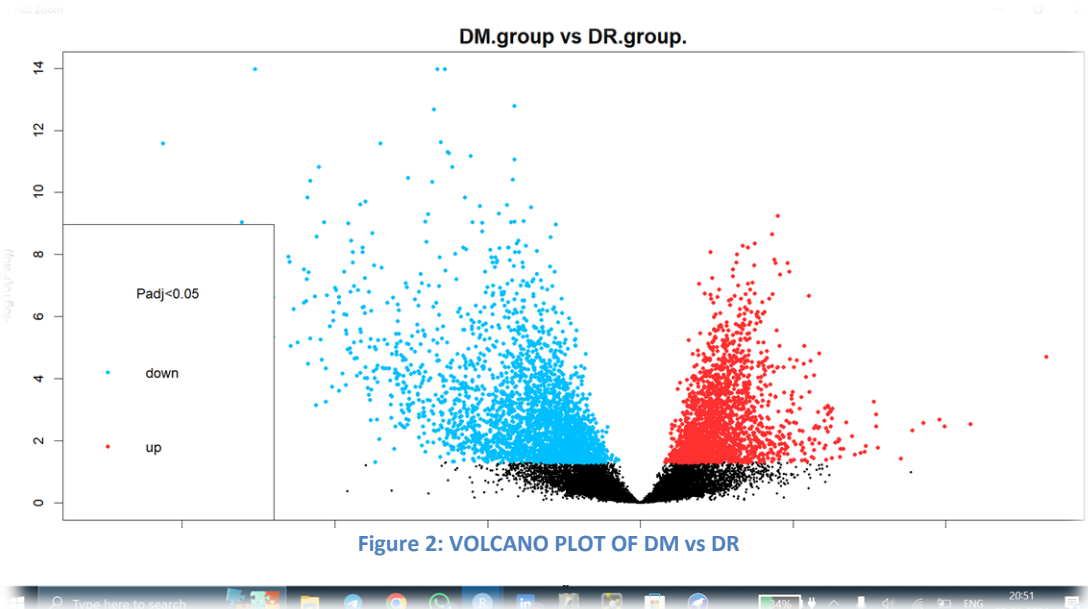

Figure 2: VOLCANO PLOT OF DM vs DR

Interpretation of Volcano Plots:

- X-axis: It represents the log2 fold change in gene expression between the two groups compared, DM and DR. Positive values indicate up regulation in one group, while negative values indicate down regulation.
- Y-axis: It represents the negative log of 10 adjusted p-values. Lower values on the y-axis indicate higher statistical significance, meaning the observed difference in expression is less likely due to chance.
- Dots: Each dot represents a gene. The color typically reflects the statistical significance (p-value) or the fold change.

  - Red: It represents genes with statistically significant up regulation (low p-value and positive fold change).
  - Blue: It represents genes with statistically significant down regulation (low p-value and negative fold change).
  - Black: It represents genes that are not statistically significant (high p-value), regardless of their fold change.

The horizontal and vertical lines define the significance cutoffs (e.g., p-value threshold and fold change threshold). Genes above the p-value line and beyond the fold change lines are typically considered differentially expressed. The red and blue dots represent the most interesting genes, showing statistically significant changes in expression between DM and DR groups. So from this volcano plot we conclude that some of the genes were differentially expressed when compared between two groups DM and DR.

2.4 Drug Target Prioritization:

2.4.1 Functional Analysis:

Functional analysis is the process of understanding the biological roles and pathways associated with genes identified as differentially expressed in the experiment. This step involves using bioinformatics tools and databases to associate DEGs with specific functions, pathways, and biological processes. KEGG Pathway this database focuses on mapping genes to

known metabolic, signaling, and regulatory pathways. By identifying enriched pathways among DEGs, you can gain insights into the overall functional consequences of the observed gene expression changes. Similar to KEGG pathway, Reactome pathway, IPA are also used for pathway and functional analysis.

Functional analysis helps you move beyond individual genes and understand their broader roles in biological processes. It can highlight genes with potential roles in disease development, drug targets, or other areas of interest. Understanding the enriched functions and pathways can guide further research questions and experiments.

DAVID:

DAVID, which stands for Database for Annotation, Visualization and Integrated Discovery, is a powerful bioinformatics resource system widely used in the field of functional analysis. It helps to understand the functions of genes in your list by identifying enriched Gene Ontology (GO) terms, KEGG pathways, and other functional categories.

For functional analysis I used DAVID because it allows combining different analyses and data sources to gain deeper insights into your genes of interest. This includes:

- Clustering
- Similarity search
- Enrichment analysis

After got results from R analysis totally 250genes were obtained which was then further moved to DAVID database. Giving gene symbols and giving identifier (Official Gene Symbol) then selected species (Homo sapiens) and then submit list. After analysis I got several gene ontology and KEGG pathway results. To do functional analysis I chose one KEGG pathway and in that pathway there were four genes from my gene list. Those four genes were selected for further network analysis.

2.4.2 Network Analysis:

Network analysis is a broad term encompassing various techniques for studying the relationships between entities. These entities can be anything from people in a social network to components in an electrical circuit.

STRING:

The STRING database, standing for Search Tool for the Retrieval of Interacting Genes/Proteins, is a valuable resource in the field of molecular biology. It serves as a biological database and web resource dedicated to known and predicted protein-protein interactions.

From DAVID analysis I have short listed 4 genes (TRRAP, SRCAP, EP400 and I NO80D). 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. All four genes show high interaction rate which makes them suitable drug targets option.

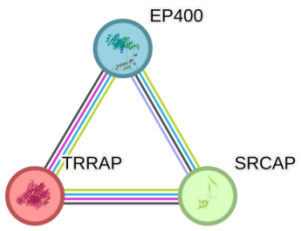

Figure 3 : Network obtained from STRING database

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 TRRAP, SRCAP, INO80D, EP400 genes had the most druggable potential.

3. Acknowledgment:

"No duty is more urgent than that of returning thanks." - James Allen

Firstly, I am deeply grateful to Bversity team for providing me with the opportunity to embark on this project and for nurturing my academic journey. I owe a special thanks to CEO of Bversity 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. Their insightful suggestions, constructive feedback, and unwavering support have been instrumental in shaping the direction of this project and bringing it to fruition. I am also thankful to the Bversity 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 and resources throughout the project. Their willingness to answer my questions and offer help has been greatly appreciated. Special thanks to the reviewers of my report, Mrs. M.Bhuvaneshwari, Dr. S.Vinoth 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. Finally, I want to thank my friends for their encouragement and support throughout this project. Their understanding and belief in me have been a source of strength and motivation. This project has been a valuable learning experience, and I am grateful to everyone who has contributed to its success.

4. Results and Communication:

4.1 Findings:

S.No.	Gene	Role in Diabetic Retinopathy	Susceptibility as Drug target for DR
01	TRRAP	Involved in chromatin remodeling and transcriptional regulation. - May contribute to increased vascular permeability and inflammation. - Plays a role in cell proliferation and migration.	Potential target due to its role in key DR pathways. - Limited research specifically on TRRAP as a DR target.
02	EP400	Histone acetyltransferase, regulates gene expression. - May contribute to increased oxidative stress and inflammatory response. - Involved in angiogenesis and vascular dysfunction.	Potential target due to its involvement in multiple DR pathways. - Some studies suggest EP400 inhibitors could protect against DR, but further research needed.
03	SRCAP	Plays a role in RNA splicing and nuclear-cytoplasmic transport. - Limited information on its specific role in DR.	Uncertain due to limited research on its role in DR. - More research needed to understand its potential as a target.
04	INO80D	Involved in DNA repair and chromatin remodeling. - May contribute to retinal cell death and dysfunction. - Limited information on its role in DR.	Uncertain due to limited research on its role in DR. - More research needed to understand its potential as a target.

4.2 Limitations:

While DEG analysis holds promise in uncovering potential therapeutic targets for diabetic retinopathy (DR), it's crucial to acknowledge several limitations:

1. Data Quality and Variability:

Different studies utilize diverse platforms, methodologies, and sample sizes, leading to inconsistent results and complicating comparisons. Variation in RNA extraction, library preparation, and sequencing protocols can introduce unwanted noise and bias. Identifying DEGs doesn't directly translate to understanding their functional role in DR development or progression.

2. Biological Complexity:

Static snapshot: DEG analysis provides a static picture at a specific time point, neglecting dynamic changes in gene expression over time in DR.

Indirect effects: Altered gene expression may be a consequence of upstream signaling pathways, making it challenging to pinpoint true drivers of disease.

Cell-type specificity: Bulk tissue analysis masks unique expression patterns in different cell types, potentially missing crucial DR-associated changes.

3. Drug Target Identification limitations:

Correlation isn't causation just because a gene is differentially expressed doesn't guarantee it's a viable drug target. Functional validation and target specificity assessments are crucial. Targeting a DEG might have unintended consequences on other processes, posing potential safety concerns. Even promising targets can face hurdles in translation to effective therapeutic agents due to technical and regulatory challenges.

5. Conclusion:

This research, utilizing DEG analysis on a dataset from NCBI and focusing on the genes TRRAP, SRCAP, INO80D, and EP400, has yielded promising findings with significant implications for future research and drug development in diabetic retinopathy (DR).

Key Findings:

This analysis suggests that TRRAP, SRCAP, INO80D, and EP400 are differentially expressed in DR compared to healthy retinas, potentially playing a role in the disease process. Understanding of DR pathways by targeting these genes, we may gain valuable insights into the underlying molecular mechanisms of DR, leading to a more comprehensive understanding of the disease. These findings pave the way for further research and development of novel therapeutic strategies specifically targeting these genes or the pathways they regulate.

Implications for Future Research:

While DEG analysis suggests their potential involvement, further studies are crucial to validate the functional roles of these genes in DR pathogenesis. This could involve in vitro and in vivo experiments to understand their specific contributions to the disease. Elucidating the precise mechanisms by which these genes contribute to DR will be essential for designing effective drugs. Drug target validation- Evaluating the suitability of these genes as drug targets is critical. This includes assessing their druggability, specificity, and potential off-target effects.

Implications for Drug Development:

Development of targeted therapies: If validated, these genes could serve as targets for developing new drugs specifically designed to treat DR. This could lead to more effective and personalized treatment options for patients. Understanding the roles of these genes could inform the design of more focused and efficient clinical trials for new DR therapies.

Overall, my research using DEG analysis has made valuable contributions to identifying potential drug targets for DR. Further research and development based on these findings hold great promise for improving the lives of patients suffering from this sight-threatening condition.