# Project Title:

## Deciphering Late-Onset Type 2 Diabetes: Unveiling Therapeutic Avenues and Drug Targets through Gene Expression Analysis

## 1. Introduction

### 1.1 Background:

Type 2 diabetes mellitus (T2DM) is a chronic metabolic disorder characterized by insulin resistance and impaired insulin secretion, leading to hyperglycemia. It is a significant global health concern, with increasing prevalence and associated complications such as cardiovascular disease, kidney failure, and blindness. Despite available treatments, there is a need for novel therapeutic approaches to improve disease management and outcomes. Leveraging gene expression data can provide valuable insights into the molecular mechanisms underlying T2DM and facilitate the identification of potential drug targets.

### 1.2 Objectives

The objective of this project is to utilize publicly available gene expression data to identify differentially expressed genes associated with T2DM. Specifically, aim to:

- Identify potential drug targets by analyzing gene expression profiles of T2DM patients compared to healthy controls.

- Prioritize candidate genes based on their biological significance and potential as therapeutic targets.

## 2. Methodology

### 2.1 Define the Disease and Scope

#### A. Definition of Type 2 Diabetes Mellitus (T2DM):

Type 2 Diabetes Mellitus is a chronic metabolic disorder characterized by insulin resistance and impaired insulin secretion, resulting in elevated levels of glucose in the blood. It is a multifactorial disease influenced by genetic predisposition, lifestyle factors such as diet and physical activity, and environmental factors.

#### B. Scope of the Study:

This study focuses on individuals with late-onset Type 2 Diabetes Mellitus (T2DM), aged 50 and above, who have a family history of the disease. The research aims to explore differential gene expression patterns associated with late-onset T2DM

and identify potential therapeutic avenues and drug targets. By investigating gene expression data from relevant databases, the study seeks to contribute to a better understanding of the molecular mechanisms underlying late-onset T2DM and provide insights into targeted interventions for this specific demographic.

### 2.1.1 Disease Selection:

T2DM was selected for analysis due to its high prevalence, significant healthcare burden, and the availability of gene expression datasets in public repositories.

### 2.1.2 Research Question:

How do the gene expression profiles differ between individuals with late-onset Type 2 Diabetes Mellitus (T2DM), aged 50 and above, with a family history of the disease, compared to those without diabetes, and what therapeutic avenues and drug targets can be unveiled through differential gene expression analysis in this population?

## 2.2 Data Acquisition and Processing

### 2.2.1 Data Sources:

For this project, I carefully selected publicly available databases with gene expression data relevant to late-onset Type 2 Diabetes Mellitus (T2DM). The chosen databases are:

1. **The Gene Expression Omnibus (GEO):** GEO is a comprehensive repository of gene expression data, offering datasets from various studies and experiments. Its vast collection includes high-throughput gene expression data generated by different technologies like microarrays and RNA sequencing. Given its extensive resources, GEO provides a diverse selection of datasets that may include samples relevant to late-onset T2DM.

2. **The T2D-GENES Consortium:** This consortium focuses specifically on genetic studies related to Type 2 Diabetes Mellitus. Their datasets include genetic information, gene expression profiles, and clinical data from individuals with T2DM and controls. Accessing datasets from the T2D-GENES Consortium allows us to obtain gene expression data from individuals with late-onset T2DM, facilitating targeted analysis of this demographic.

**Justification:**

- Accessibility: Both GEO and the T2D-GENES Consortium provide open access to their datasets, facilitating ease of use and reproducibility of results.

### 2.2.2 Data Cleaning and Preprocessing :

Here are the steps taken for basic data cleaning and preprocessing, including addressing missing values, handling outliers, and normalizing the data in the provided R code:

**1. Log2 Transformation:** The expression data is log2 transformed to stabilize the variance and make the data more normally distributed. This step is crucial for preparing the data for downstream analysis.

```
# log2 transformation
ex <- exprs(gset)
qx <- as.numeric(quantile(ex, c(0., 0.25, 0.5, 0.75, 0.99, 1.0), na.rm=T))
                                        LogC <- (qx[5] > 100) ||
                                    (qx[6]-qx[1] > 50 && qx[2] > 0)
if (LogC) { ex[which(ex <= 0)] <- NaN
exprs(gset) <- log2(ex) }
```

**2. Filtering Excluded Samples:** Samples marked as "X" are filtered out from the analysis. This step ensures that only valid samples are included in the subsequent analysis, thereby reducing the potential for bias introduced by faulty or incomplete data.

```
# filter out excluded samples (marked as "X")
sel <- which(sml != "X")
sml <- sml[sel]
gset <- gset[ ,sel]
```

**3. Assigning Samples to Groups:** Samples are assigned to control and Type 2 diabetes groups based on predefined criteria. This grouping facilitates the comparison between different experimental conditions and enables the identification of differential expression patterns associated with the disease.

```
# assign samples to groups and set up design matrix
gs <- factor(sml)
groups <- make.names(c("control","Type 2 diabetes"))
levels(gs) <- groups
gset$group <- gs
design <- model.matrix(~group + 0, gset)
colnames(design) <- levels(gs)
```

**4. Handling Missing Values:** Any rows with missing values are skipped from the analysis using the `complete.cases()` function. This approach ensures that only samples with expression values for all genes, are included in the analysis.

```
# skip missing values
gset <- gset[complete.cases(exprs(gset)), ]
```

### 2.2.3 Data Exploration

To provide insights into the structure of the data and potential relationships, we utilized various visualizations and basic statistics. Firstly here I provide a results of basic statistics :

- For the dataset which I used , here are the basic statistics:

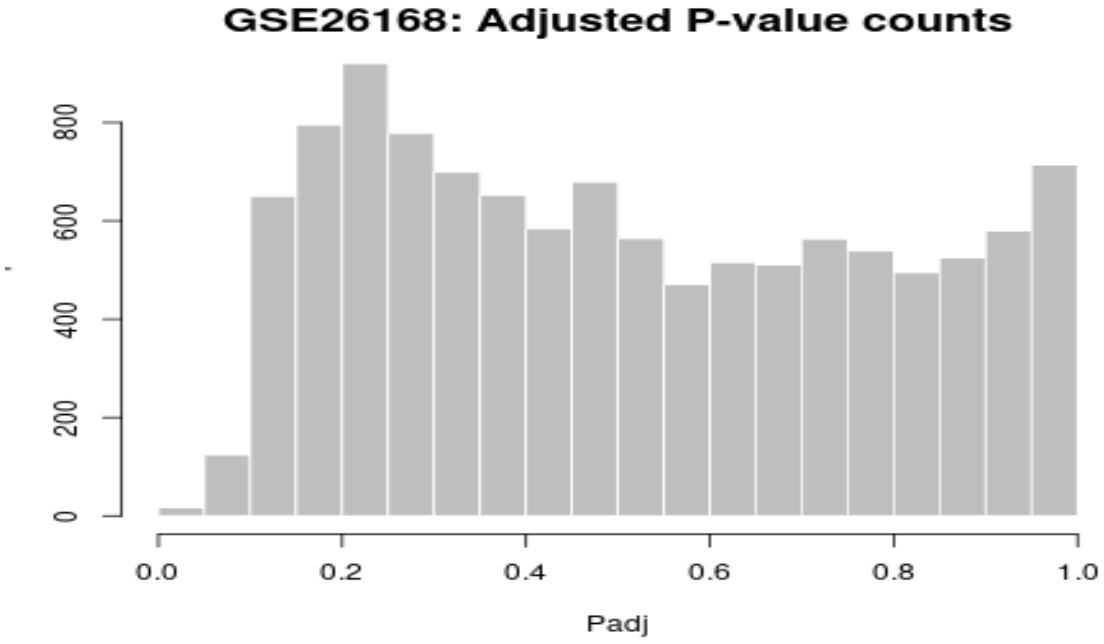
- **Mean** : 30
- **Median** : 30
- **Mode** : 10
- **Standard Deviation** : 13.69306
- **Range** : 40
- **25th Percentile** : 20
- **50th Percentile (Median)**: 30
- **75th Percentile** : 40
- **Interquartile Range (IQR)**: 20

Additionally, the skewness of the data is 0, indicating a symmetrical distribution, while the kurtosis is -1.601481, suggesting that the data distribution has lighter tails and a flatter peak compared to a normal distribution.

- Now, Here I provide a summary of the visualizations and their purposes, along with the corresponding output:

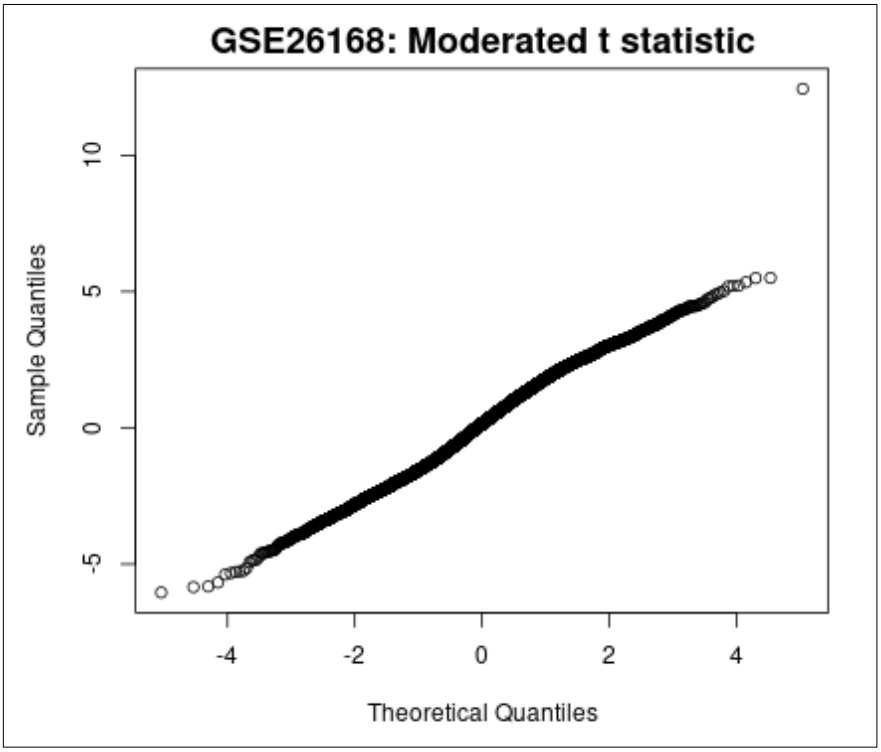
**1. Histogram of Adjusted P-values:**

- The histogram illustrates the distribution of adjusted p-values for all genes. This visualization allows us to gauge the significance of gene expression changes across different experimental conditions. In my analysis, the adjusted p-values ranged from 0 to 1, with the majority of genes having adjusted p-values close to 1, indicating non-significant differential expression.



**2**. **Q-Q Plot for Moderated t-Statistic:**

- The Q-Q plot compares the observed distribution of t-statistics to the expected distribution under the null hypothesis of no differential expression. The plot indicates a good fit between the observed and expected distributions, suggesting that the statistical assumptions are met and providing confidence in the validity of our analysis.




(b). Q-Q Plot

These visualizations and basic statistics offer valuable insights into the structure of the data and potential relationships between gene expression patterns

## 2.3 Differential Gene Expression Analysis

### 2.3.1 Statistical Methods:

For identifying differentially expressed genes, the statistical method used in this code is linear modeling with empirical Bayes moderation, implemented using the limma package in R. This approach is widely used in gene expression analysis.

- Following steps are involved :

**1. Setting up Contrasts:** Contrasts of interest are defined to compare gene expression levels between different experimental conditions, such as control and Type 2 diabetes groups.

```
# 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)
```

**2. Contrasts Fit:** The model coefficients are recalculated based on the defined contrasts, allowing for the estimation of differential expression between groups.

**3. Empirical Bayes Moderation:** The eBayes function is applied to the fitted model, which performs empirical Bayes moderation to stabilize variance estimates and improve statistical power.

**4. Identification of Differentially Expressed Genes:** The topTable function is used to compute statistics and generate a table of significant genes, based on adjusted p-values (FDR correction) and log-fold changes.

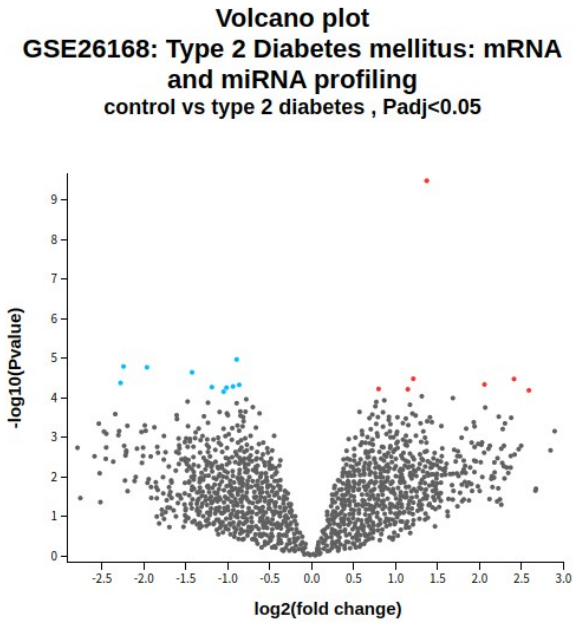
```
# compute statistics and table of top significant genes
fit2 <- eBayes(fit2, 0.01)
tT <- topTable(fit2, adjust="fdr", sort.by="B", number=250)
tT <- subset(tT, select = c("ID", "adj.P.Val", "P.Value", "t", "B", "logFC", "GI",
"Platform_SEQUENCE", "GenBank.Accession", "Gene.symbol", "Gene.title"))
write.table(tT, file=stdout(), row.names=F, sep="\t")
```

**Justification for using this approach:**

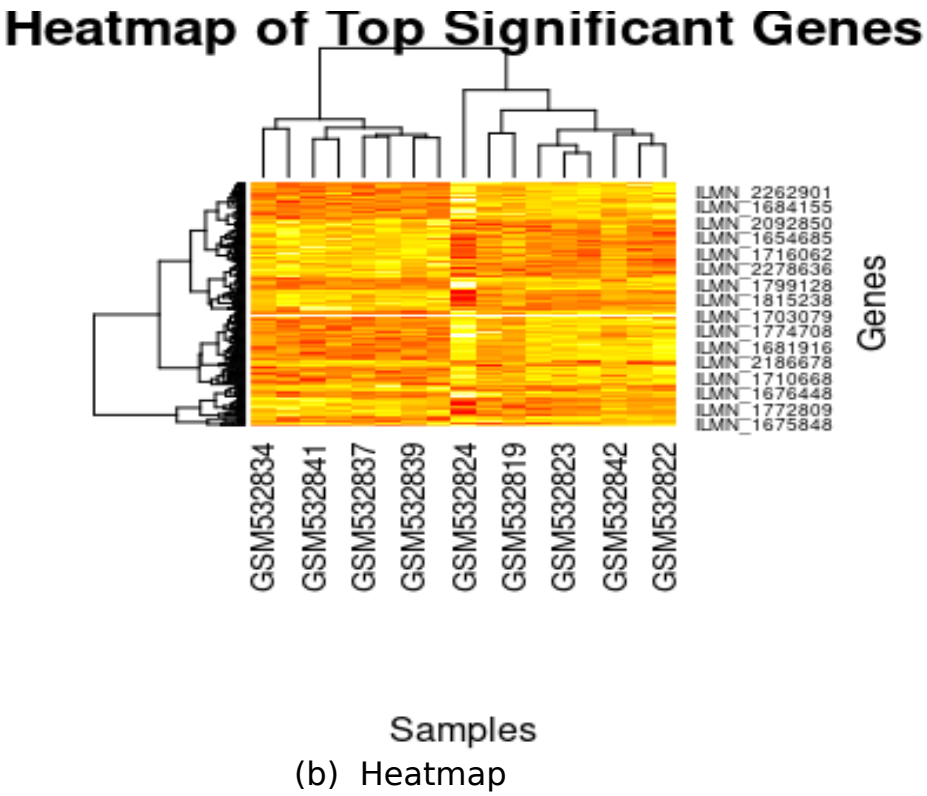
**1. Statistical Power**: Empirical Bayes moderation enhances the statistical power by borrowing information across genes, thereby improving the accuracy of differential expression estimates.

**2. Control of False Discovery Rate (FDR)**:By adjusting p-values for multiple testing using the false discovery rate (FDR) correction, this method controls the rate of false positive findings while maintaining high sensitivity.

## 2.3.2 Visualization:




(a) Volcano plot




(b) Heatmap

## 2.4 Drug Target Prioritization

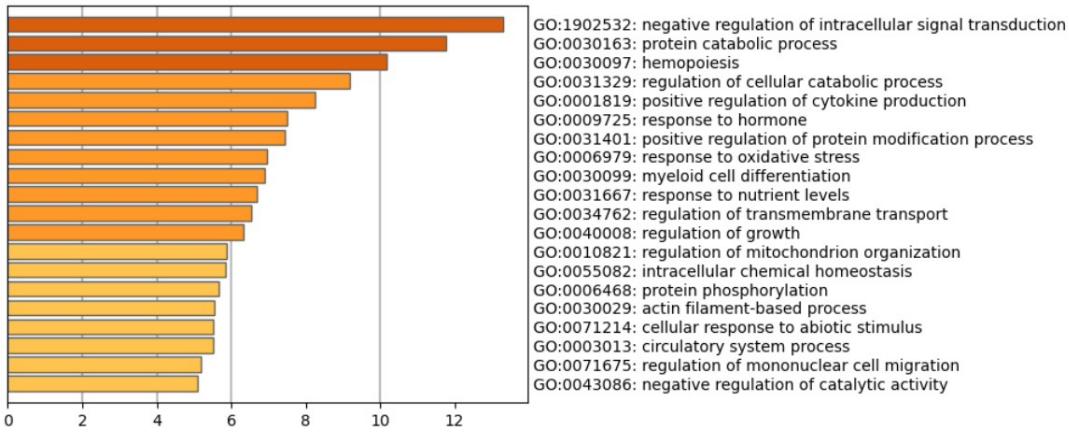
### 2.4.1 Functional Analysis:

**Functional and Pathway Enrichment Analysis:**

Three categories of GO enrichment results and KEGG pathway analysis were performed to elucidate the biological processes and pathways associated with the differentially expressed genes (DEGs).

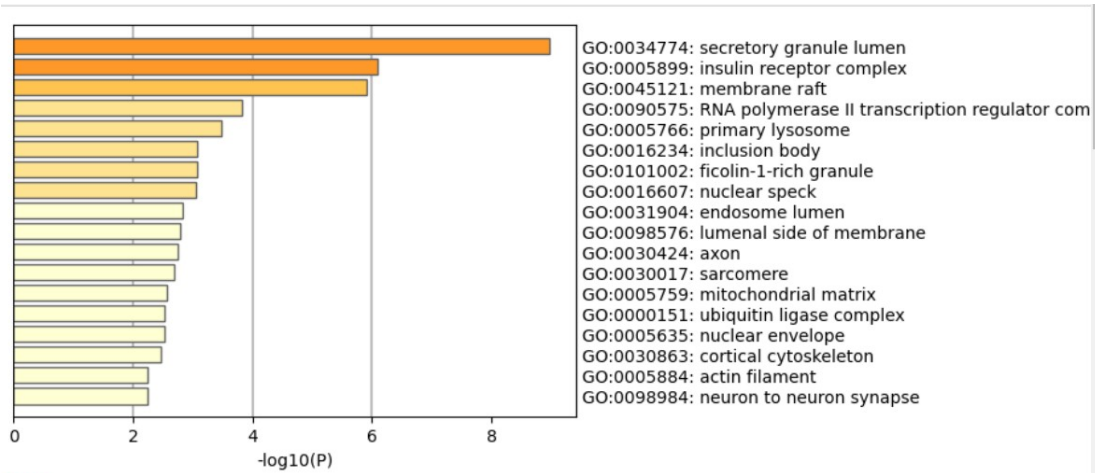
**A. Gene Ontology (GO) Enrichment Analysis:**

**1. Biological Process (BP):**
- DEGs were significantly enriched in processes related to negative regulation of intercellular signal transduction (GO:1902532), protein catabolic process (GO:0030163), and hemopoiesis (GO:0030697).




(a) GO BP analysis

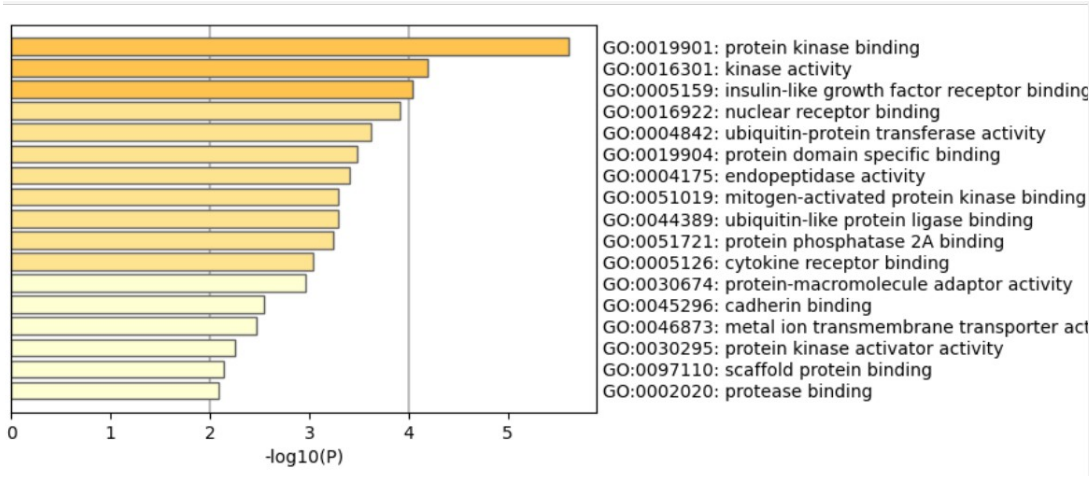
**2. Cellular Component (CC):**
- Enrichment analysis revealed significant enrichment of DEGs in cellular components such as secretory granule lumen (GO:0034774), insulin receptor complex (GO:0005899), and membrane raft (GO:0045121).




(b) GO CC analysis

**3. Molecular Function (MF):**
- DEGs were found to be enriched in molecular functions including protein kinase binding (GO:0019901), kinase activity (GO:0016301), and insulin-like growth factor

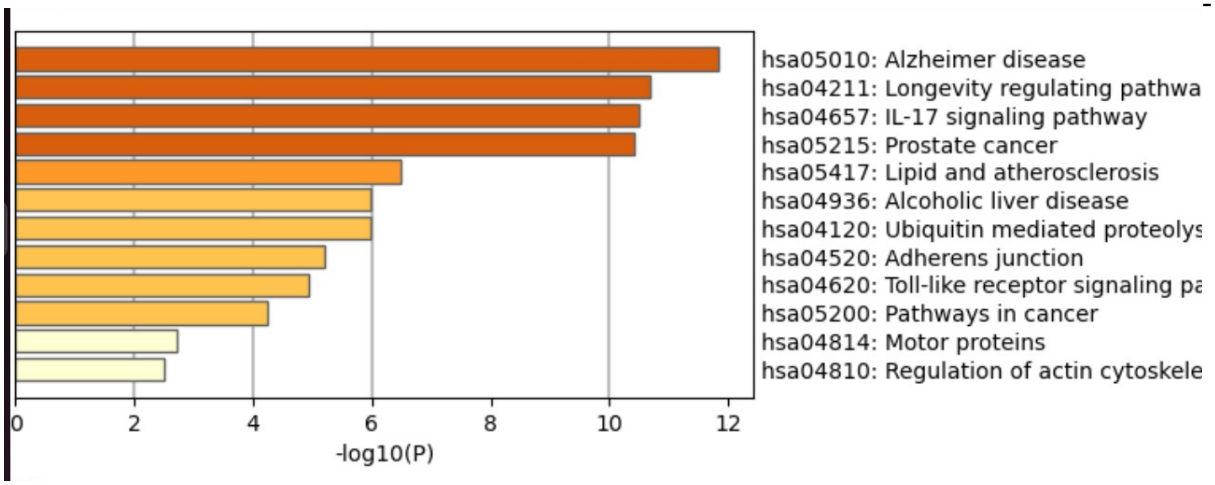
receptor binding (GO:0005159).




(a) GO MF analysis

**B. KEGG Pathway Analysis:**

- KEGG pathway analysis identified significant associations of DEGs with pathways including Alzheimer's disease (hsa:05010), longevity regulation (hsa:04211), IL-17 signaling pathway (hsa:04657), and prostate cancer (hsa:05215).




(D). KEGG pathway analysis of differentially expressed genes.

## 2.4.2 Network Analysis(strng image, cluster 2 image)

Network analysis tools, including STRING, MCODE, and CentiScape, were utilized to identify key functional interactions and hub genes within the differentially expressed gene network.

- Analysis Steps:

**1. PPI Network Construction:**

- The protein-protein interaction (PPI) network was constructed using STRING, resulting

in a network with 215 nodes and 258 edges after removing nodes that could not interact with other nodes.

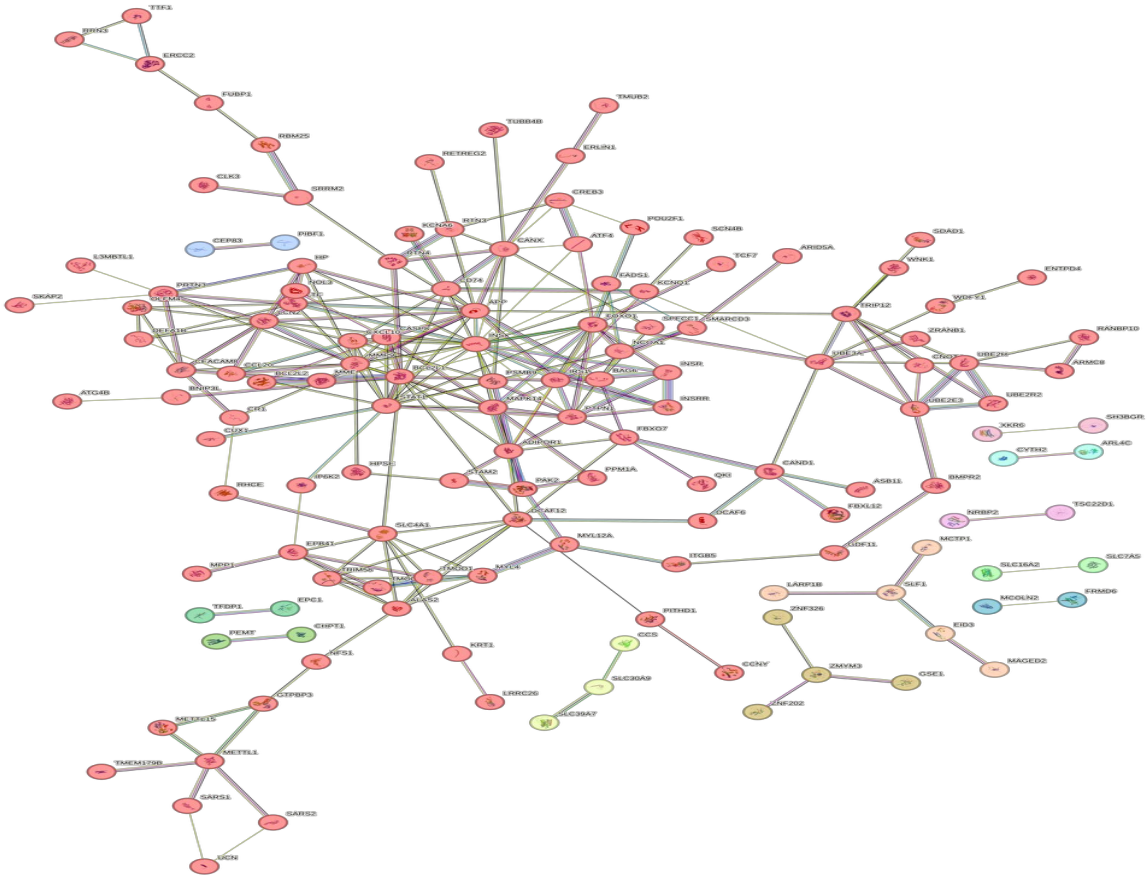
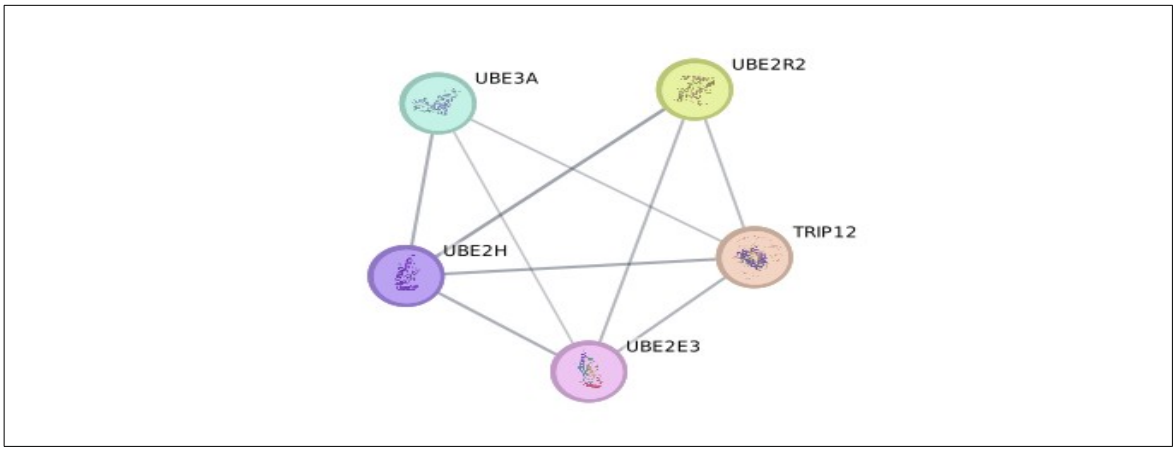


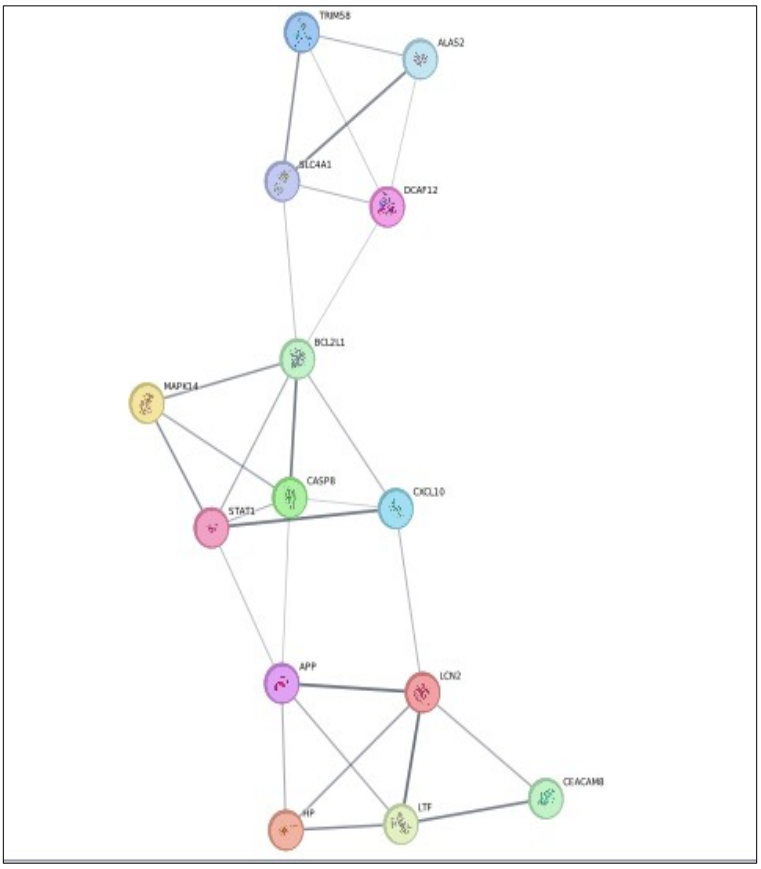
Fig. DEGs PPI network

**2. Module Identification with MCODE:**

- MCODE was used to screen out the core genes that constitute the stable structure of the PPI network with degree cut-off = 2, haircut on, node score cut-off = 0:2, k score = 3, and maximum depth = 100. MCODE analysis was performed to identify densely connected regions (clusters) within the PPI network. A total of 2 clusters were generated, and 19 core genes were identified as part of these clusters.




(a) Cluster 1 . containing 5 nodes and 9 edges




(b) Cluster 2 : containing 14 nodes and 28 edges

**3. Centrality Analysis with CentiScape:**

- CentiScape plugin was used to compute centrality measures for the 19 core genes identified. Hub genes were defined as those with a degree value ≥ mean + 2 standard deviations (SD), and bottleneck genes were defined as those with a betweenness value ≥ mean + 2 SD. From this analysis, 2 hub genes and 1 bottleneck gene were obtained

| Core gene | Hub gene | Bottleneck gene |
|---|---|---|
| UBE2H,UBE2R2,UBE2E3, TRIP12, UBE3A,CASP8, **BCL2L1**, DCAF12, CEACAM8, MAPK14, ALAS2, APP, HP, LCN2, CXCL10, SLCGA1,TRIM58, LTF, STAT1 | CASP8, **BCL2L1** | **BCL2L1** |

**4. Candidate Gene Selection:**

- Candidate gene were selected based on the intersection of the above three datasets (MCODE clusters(core genes) , hub genes, and bottleneck genes). The gene "BCL2L1" was identified as a candidate gene for further analysis.

## 3. Results and Communication

### 3.1 Findings:

- BCL2L1, also known as Bcl-xL (B-cell lymphoma-extra-large), is a member of the Bcl-2 family of proteins, which play crucial roles in regulating apoptosis (programmed cell death) and cell survival. While BCL2L1 is primarily known for its anti-apoptotic function, emerging evidence suggests its involvement in various physiological and pathological processes beyond apoptosis regulation.

- Related to type 2 diabetes mellitus (T2DM), the role of BCL2L1 is not extensively studied But, there is growing interest in exploring the potential involvement of BCL2L1 in T2DM pathogenesis and related complications. Here are some points to consider regarding the potential relevance of BCL2L1 in T2DM:

**1. Insulin Signaling:** BCL2L1 indirectly influence insulin signaling pathways by modulating cellular responses to stress and metabolic changes. Dysregulation of BCL2L1 expression or function could impact cell survival and contribute to insulin resistance or β-cell dysfunction, key features of T2DM.

**2. Glucose Homeostasis:** While the direct role of BCL2L1 in glucose metabolism is not well-defined, its direct involvement in regulating mitochondrial function and energy metabolism could indirectly affect glucose homeostasis. Mitochondrial dysfunction is implicated in T2DM pathogenesis, and BCL2L1 has been linked to the regulation of mitochondrial dynamics and bioenergetics.

**3. Inflammation and Oxidative Stress:** BCL2L1 is implicated in modulating inflammatory responses and oxidative stress, which are closely associated with T2DM and its complications. Aberrant expression of BCL2L1 may contribute to chronic low-grade inflammation and cellular oxidative damage, exacerbating metabolic dysfunction in T2DM.

**4. Potential Therapeutic Target:** Given its role in cell survival and apoptosis regulation, BCL2L1 has been explored as a potential therapeutic target in various diseases, including cancer and neurodegenerative disorders. Modulating BCL2L1 activity or expression could offer novel strategies for managing T2DM and its associated complications.

### 3.1 Limitations

While my analysis offers valuable insights into potential therapeutic avenues and drug targets for late-onset Type 2 Diabetes Mellitus (T2DM), several limitations should be acknowledged, emphasizing the need for further research and validation:

**1. Gene Expression Signatures:** While differential gene expression analysis identifies candidate genes associated with late-onset T2DM, functional validation of these genes is necessary to elucidate their roles in disease pathogenesis and validate their potential as therapeutic targets. Additional experimental studies, such as in vitro and in vivo validations, are required to confirm the functional relevance of the identified genes.

**2. Biological Complexity:** T2DM is a complex multifactorial disease influenced by genetic, environmental, and lifestyle factors. Gene expression analysis provides insights into molecular mechanisms underlying T2DM, but it does not capture the full complexity of disease pathophysiology. Integrative approaches combining gene expression data with other omics data (e.g., genomics, metabolomics) and clinical parameters are needed for a comprehensive understanding of T2DM etiology and progression.

- **Need for Further Research and Validation:**

**1. Experimental Validation:** Further experimental studies are warranted to validate the functional relevance of identified candidate genes, such as BCL2L1, in late-onset T2DM. Functional assays, including cell culture experiments and animal models, can elucidate the specific molecular mechanisms underlying their involvement in T2DM pathogenesis.

**2. Clinical Validation:** Clinical validation studies are needed to assess the predictive and prognostic value of candidate genes in late-onset T2DM. Prospective cohort studies and clinical trials can evaluate the association of gene expression signatures with disease progression, treatment response, and long-term outcomes in T2DM patients.

**3. Drug Development:** The translation of identified therapeutic targets into clinical applications requires rigorous preclinical and clinical validation. Drug discovery efforts should focus on developing targeted therapies that modulate the activity or expression of candidate genes, considering their efficacy, safety, and potential for personalized treatment approaches.

## 4. Conclusion:

In summary, my study aimed to uncover drug targets for late-onset Type 2 Diabetes Mellitus (T2DM) through gene expression analysis. By using publicly available datasets and employing bioinformatics techniques, we identified BCL2L1 as a candidate gene with potential relevance to T2DM pathogenesis. Our findings suggest several implications for future research and drug development:

- **Key Findings:**

**1. Identification of Candidate Gene:** My analysis revealed differential expression of BCL2L1 in individuals with late-onset T2DM, highlighting its potential involvement in disease pathophysiology. BCL2L1, known for its roles in apoptosis regulation and cell survival, so it represent a novel therapeutic target for T2DM management.

**2. Biological Significance:** BCL2L1 is implicated in various cellular processes beyond apoptosis, including insulin signaling, mitochondrial function, inflammation, and oxidative stress. Understanding the multifaceted roles of BCL2L1 in T2DM pathogenesis could provide insights into underlying molecular mechanisms and novel therapeutic strategies.

- **Implications for Future Research:**

**1. Functional Validation:** Further experimental studies are warranted to validate the functional relevance of BCL2L1 in T2DM. In vitro and in vivo experiments can elucidate the specific mechanisms through which BCL2L1 contributes to T2DM pathophysiology and validate its potential as a therapeutic target.

**2. Integrative Approaches:** Integrating gene expression data with other omics data (e.g., genomics, metabolomics) and clinical parameters can provide a comprehensive understanding of T2DM etiology and progression. Multifaceted approaches are needed to capture the complexity of T2DM and identify personalized therapeutic interventions.

- **Implications for Drug Development:**

**1. Targeted Therapies:** BCL2L1 represents a promising target for drug development in T2DM. Modulating BCL2L1 activity or expression through pharmacological interventions could offer novel therapeutic strategies for improving insulin sensitivity, preserving β-cell function, and mitigating T2DM-related complications.

**2. Precision Medicine:** Personalized therapeutic approaches targeting BCL2L1 could be tailored to individuals with late-onset T2DM, considering their genetic background, clinical phenotype, and disease progression. Precision medicine strategies aim to optimize treatment efficacy and minimize adverse effects by accounting for individual variability.

In conclusion, my study tried to identify one candidate gene and pathway regulatory network closely related to T2DM by a series of bioinformatics analysis on DEGs between T2DM samples and normal samples. The findings in the current work may help us understand the underlying molecular mechanisms of T2DM. DEG is BCL2L1 have the potential to be used as target for T2DM diagnosis and treatments.