# Peripheral Blood Gene Expression in Diabetic Retinopathy Patients

Jaden Thomas

2024-11-12

## Introduction

Diabetes Mellitus is a chronic disease that affects more than 830 million people worldwide (“Diabetes” 2024). It is characterized by elevated blood glucose levels, either because of an inability to produce insulin or ineffective use of produced insulin. The three most common classifications of diabetes are type 1, type 2 and gestational diabetes (“Diabetes Basics Diabetes CDC”). Diabetes is very heterogeneous in its sub classifications in both mechanisms of action and health outcomes (Udler et al. 2018). Diabetes is a comorbidity that increases risk for many other chronic diseases such as heart disease and cancer. One outcome of mismanaged blood glucose levels in diabetic individuals is diabetic neuropathy. Neuropathy is the result of damage to the peripheral and autonomic nervous system due to exposure to high glucose levels (Feldman et al. 2019). This can result in nerve damage in the peripheral parts of the body, such as the hands, feet, and eyes. Diabetic Retinopathy is the subset of diabetic neuropathy that affects the eyes and damages blood vessels in the eyes.

The data in this analysis comes from Xiang et al. (Xiang et al. 2023), which is a study to determine the changes and related factors of blood CCN1 levels in diabetic patients. The study cohort includes 193 participants (95 female, 98 male). There are three groups in the study, control, DM (Diabetes mellitus), and DR (Diabetic retinopathy).

The participant’s mRNA was analyzed using Illumina NovaSeq 6000. The mRNA data and supplemental data were downloaded from the Gene Expression Omnibus (GSE221521). In this paper, further analysis of the DR and DM patients was conducted to determine differences in mRNA expression in the two groups and if the expression could be used to predict or separate the groups.

## Methods

The original data includes 60675 genes from leukocytes in 193 participants. After filtering for low counts, 16697 genes were used for downstream analysis.

## Pre-Processing

The RNA-Seq raw counts were filtered to exclude genes with an average of less than 10 counts per sample. A variance stabilizing transformation was then applied to the filtered counts using DESeq2 (Love et al. 2014). Data was filtered to only include participants with DM or DR to analyze differences in the two groups.

## Differential Expression

Differential expression analysis was performed on the filtered counts using a Negative Binomial model with DESeq. Comparisons between Diabetic Neuropathy and Diabetic Mellitus, and Diabetic Neuropathy and Control were analyzed. Differentially expressed genes were determined by q-value<0.05 and $|log_2 \text{ Fold Change}| > 1$. Gene ontology of differentially expressed genes was then analyzed using clusterProfiler (Yu et al. 2012) and the org.Hs.eg.db Entrez gene wide annotation for humans.

## Supervised Learning

### Lasso Regression

DR and DM participants were filtered to analyze differences in RNA-seq between the two. The variance stabilized data of those participants was scaled, and then a Lasso logistic regression was performed to predict DR, with DM as the control. A 70-30 train test split was used to train the model. 10-fold cross validation was used to determine the lambda which minimized binomial deviance. Test error, confusion matrix and AUC were used to evaluate model performance.

### Random Forest

The variance stabilized data was centered and scaled, and then a random forest classification model was fit to predict the class, of DR and DM. A 70-30 training split was used to fit the random forest. The training data was trained with n=500 trees. OOB error for the training data, confusion matrix, AUC, and test error were used to evaluate model performance. Variable importance was also analyzed using mean decrease in accuracy as the importance measure.

## Dimension Reduction

The variance stabilized data was normalized using quantile normalization. Data was filtered to include expression data from participants with DR or DM. T-SNE, PCA, and UMAP were performed to reduce the dimensions and compare components between DR and DM.

# Results

*Table 1:* Distribution of Response Variable

| Distribution of Response Variable | |
|---|---|
| Group | n |
| Control | 50 |
| DM | 74 |
| DR | 69 |

## Differential Expression

Differential expression of DR vs DM participants revealed 3829 differentially expressed genes with an adjusted p-value $\leq 0.05$. 2320 genes were upregulated in DR participants and 1509 were downregulated in DR participants. Of those 3829 differentially expressed genes, 43 had $|log_2 \text{ Fold Change}| > 1$. Some of the downregulated genes include RN7SK, AC012321.1, RPS9, CEACAM6, and OLR1. The upregulated genes include ATP5PDP4, TATDN2P2, AP005131.1, EEF1A1P16, and ACTBP2. AC097637.1is the only differentially expressed gene with a $|log_2 \text{ Fold Change}| > 2$, shown in Figure 1A. The biological processes these genes are involved include RNA splicing, ribonucleoprotein complex biogenesis, and ncRNA processing, as shown by gene ontology analysis in Figure 1.

Analysis of DR vs Control participants resulted in 5110 differentially expressed genes. Of these genes, 2850 were upregulated in DR, and 2260 were downregulated. The top 5 upregulated genes based on ascending adjusted p-value are ATP5PDP4, AC127070.4, ACTBP2, TATDN2P2, AC006059.1. The top 5 downregulated genes are RPL3P4, LINC00853, TRAV25, ADTRP, and TRAV1. AC012321.1 was the only significant gene with $|log_2 \text{ Fold Change}| > 2$. Some of the biological processes that these differentially expressed genes are involved with are very similar to those differentially expressed in DR vs DM.

The differential expression analysis gives new genes and pathways for analysis that may play a role in the development of retinopathy in diabetic patients and can help to understand the biological processes and expression in retinopathy patients compared to diabetic patients without retinopathy and non-diabetic patients.

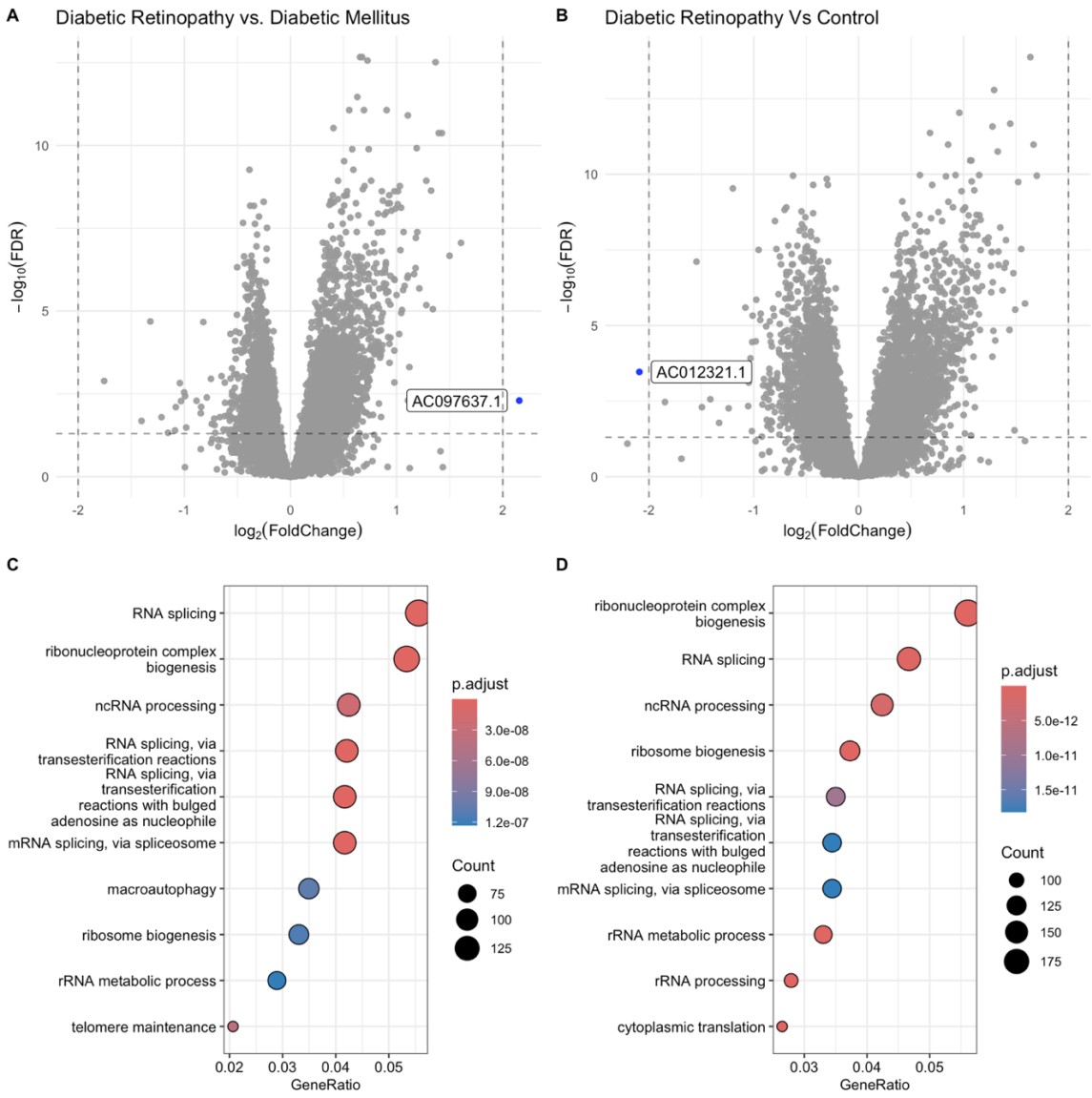


*Figure 1:* Differential expression analysis of Diabetic Retinopathy participants. (A-B) Volcano plot comparing RNA-seq expression of Diabetic Retinopathy participants with Diabetic Mellitus (A) and Control (B) participants. Labeled genes represents FDR<0.05 and absolute value of fold change > 2. (C-D) Gene ontology analysis showing biological process of significant genes from differential expression analysis for DR vs. (C) DM, and (D) Control. Benjamini-Hochberg adjustment for multiple comparisons was used for False Discovery Rate (FDR).

## Class Prediction

Lasso binomial regression and random forest models were trained to classify participant groups on the subset of participants that have DR or DM. For the models, DM was the baseline group, and DR was the experimental group. The cross-validation for the lambda parameter in the lasso model can be seen in Figure 2A. The OOB error rate for

random forest was 23.29%, while the DM error rate (10.26%) was much lower than the DR error rate. (38.24%). The final lasso model had a 32.86% test error, and random forest had a 24.29% test error. The random forest model also had a higher Area Under the Curve (AUC), with 0.7384, while lasso had 0.7012. The receiver operating characteristic (ROC) for both models can be seen in Figure 2C. The confusion matrix for lasso and random forest can be seen in Table 3 and Table 4.

*Table 2:* Top 10 Lasso Variable by Absolute Coefficient Estimate

| Gene | Lasso Beta |
|---|---|
| TFIP11 | -0.6511331 |
| AL356750.1 | 0.5298793 |
| FRG1BP | -0.3941159 |
| NGRN | -0.3665671 |
| AC093690.1 | 0.2596843 |
| TBC1D3K | 0.2574916 |
| GLDC | -0.2553448 |
| AL031595.2 | 0.2496158 |
| GOLGA8A | 0.2416086 |
| TMEM254-AS1 | 0.2216007 |

The lasso model resulted in variable selection with 35 nonzero gene coefficients, shown in Table 2. The genes with the highest absolute coefficient values were TFIP11, AL356750.1, and FRG1BP. The most important variables in the random forest model by Mean Decrease in Accuracy were MED4, PLXND1, PIEZO1, NCL, and AC073957.3 (Figure 2B).

These models both predict whether diabetic patients have retinopathy well, with >70% test accuracy with the random forest, and >65% for lasso. Random forest predicted also DM much better than lasso, with the training error being 10.26% and only misclassifying 4 actual DM, while lasso misclassified 12 actual DM.

*Table 3:* Lasso Logistic Regression Confusion Matrix

| Lasso Confusion Matrix | | |
|---|---|---|
| Predicted Group | Actual Group | Count |
| DM | DM | 23 |
| DR | DM | 12 |
| DM | DR | 11 |
| DR | DR | 24 |

*Table 4:* Random Forest Confusion Matrix

| Random Forest Confusion Matrix | | |
|---|---|---|
| Predicted Group | Actual Group | Count |
| DM | DM | 31 |
| DR | DM | 4 |
| DM | DR | 13 |
| DR | DR | 22 |

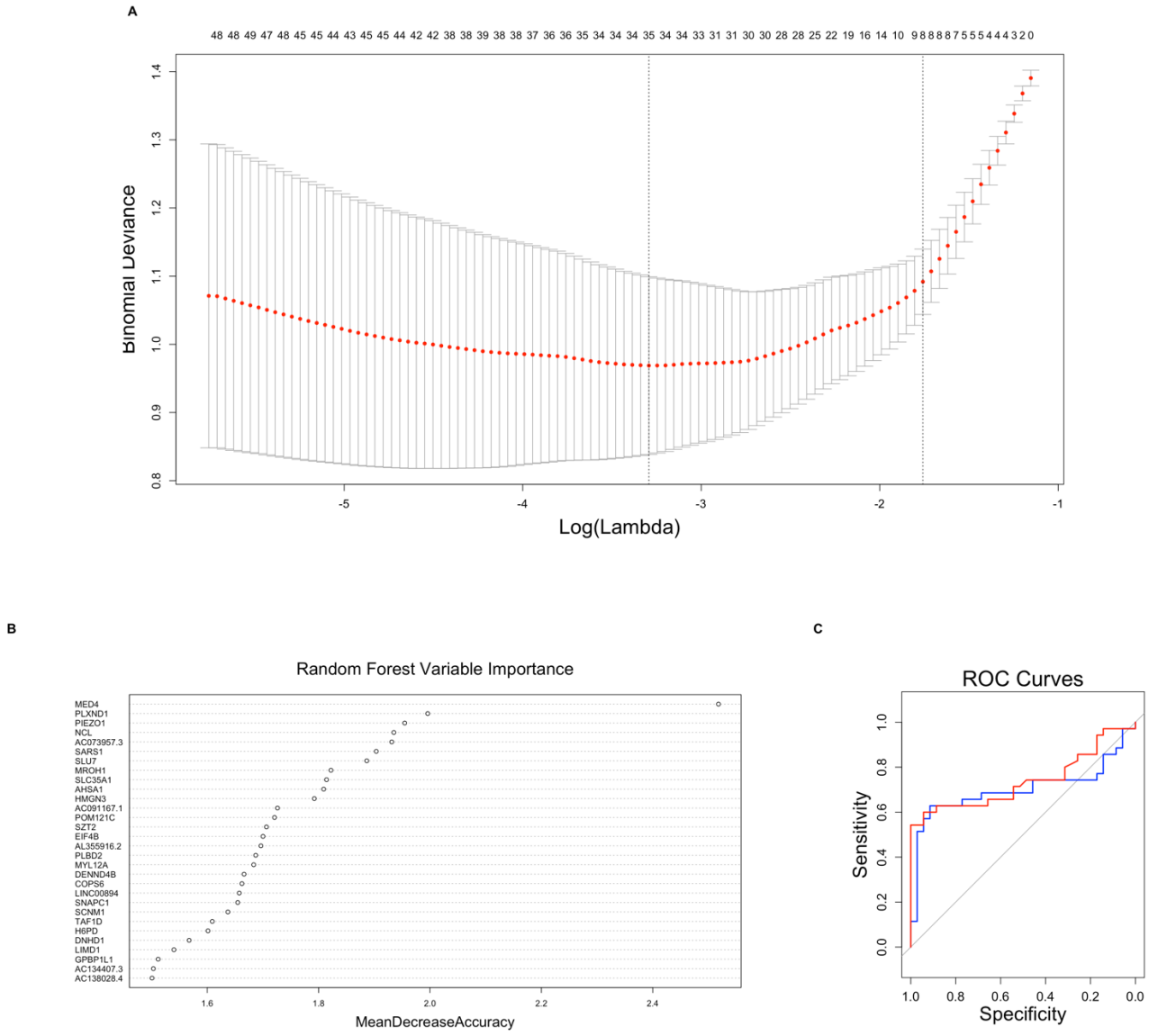


*Figure 2:* Supervised learning analysis of DR and DM patients using gene expression. (A) Cross validation for lambda lasso parameter. X axis shows the log lambda value, and Y axis shows Binomial Deviance. Lambda is chosen to minimize log lambda. (B) Variable importance plot showcasing the top 30 genes ranked by Mean Decrease in Accuracy. (C) Receiver Operating Characteristic plots of Lasso (blue), and Random Forest (red).

## Dimension Reduction

Principal component analysis (PCA) was performed on the gene expression data for dimension reduction. Reducing the data to 98 principal components (PCs) explains 90% of the data. The scree plot for principal components is shown in Figure 3A. The first 2 principal components explain 13.45% and 12.97% respectively, shown in Figure 3C. The

DR and DM groups seem to be overlapping a large amount, suggesting just 2 PCs may not be enough to distinguish between DR and DM.

T-SNE was also performed to reduce the data into two-dimensions using non-linear dimension reduction. The first 2 components of T-SNE are shown in Figure 3B. The data seems to split with DR (red) having a lower first component compared to DM (black). There is still overlap, but the groups are much more distinguishable compared to PCA. Finally, UMAP was performed to reduce the dimensions of the gene expression, shown in Figure 3D. The UMAP method seems to produce a well-defined cluster in the top right of DR patients. This may suggest there is a cluster of DR patients with very similar gene expression, and some that have expression like DM patients which did not get reduced to form the cluster.

These results suggest that dimension reduction can be used to separate DR and DM patient gene expression. T-SNE and UMAP specifically seem to both separate, and perhaps form clusters. The result dimension reduction techniques could be used for further analysis to determine why these patients cluster together, and if the components can be used to predict DR.

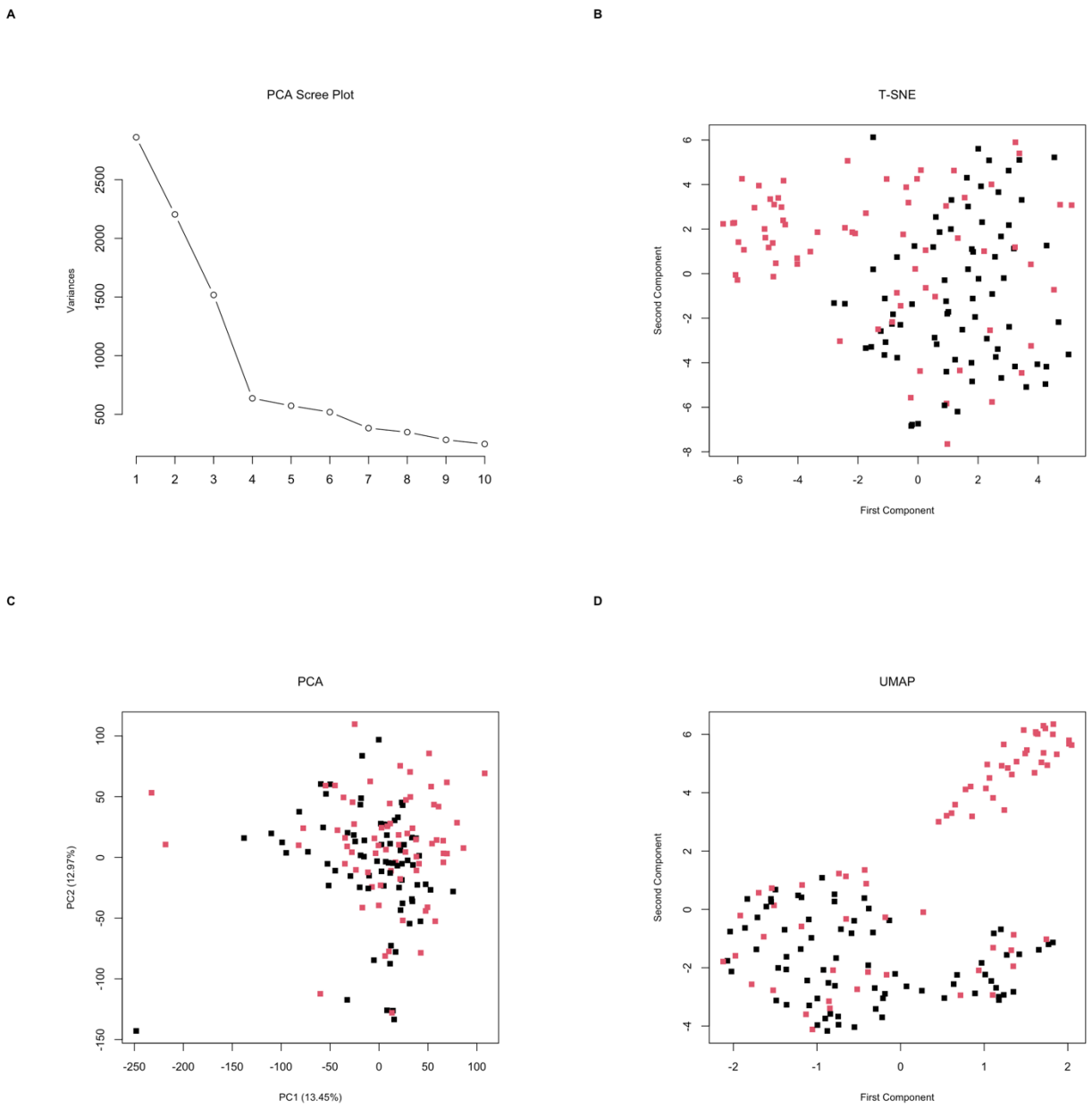


*Figure 3:* Dimension reduction results for DR and DM patients using gene expression data. (A) Scree plot showing Principal Components (PCs) on the x-axis and variance explained on the y-axis. (B) First 2 components of T-SNE dimension reduction. (C) First 2 components of PCA, with PC1 (13.45%) on the x-axis and PC2 (12.97%) on the y-axis. (D) First 2 components of UMAP. (B-D) DR shown in red and DM shown in black.

# Discussion

Analysis of mRNA-Seq data from 143 diabetic patients and 50 non-diabetic patients has shown that gene expression can be used to distinguish between diabetic neuropathy and diabetic without neuropathy and more insights for analysis of diabetic neuropathy. Differential expression analysis showed several genes that may give insight for the

differences between diabetic with retinopathy and diabetic without retinopathy and provide genes for further analysis. Supervised learning analysis also gave insight into genes that are important in directly predicting DR and DM, as well as methods for predicting DR with reasonable accuracy based on gene expression, including lasso and random forest. Also, dimension reduction analysis has shown methods that could be used to reduce gene expression of patients and separate DR and DM, especially T-SNE and UMAP. These analyses provide insight into retinopathy and future research of treatment based on differential gene expression and variable importance in prediction models, as well as prediction for patient's retinopathy status.

Future directions of analysis include further analysis of genes that were differentially expressed and or important in prediction, as well as improved predictive models that include more covariate data. Also, further analysis of dimension reduction results could result in computationally simpler methods for predicting retinopathy in patients.

# References


"Diabetes." 2024. https://www.who.int/news-room/fact-sheets/detail/diabetes.

"Diabetes Basics  Diabetes  CDC." n.d. https://www.cdc.gov/diabetes/about/index.html. Accessed November 18, 2024.

Feldman, Eva L., Brian C. Callaghan, Rodica Pop-Busui, Douglas W. Zochodne, Douglas E. Wright, David L. Bennett, Vera Bril, James W. Russell, and Vijay Viswanathan. 2019. "Diabetic Neuropathy." *Nature Reviews Disease Primers* 5 (1): 1–18. https://doi.org/10.1038/s41572-019-0092-1.

Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014). https://doi.org/10.1186/s13059-014-0550-8

Udler, Miriam S., Jaegil Kim, Marcin von Grotthuss, Sílvia Bonàs-Guarch, Joanne B. Cole, Joshua Chiou, Christopher D. Anderson on behalf of METASTROKE and the ISGC, et al. 2018. "Type 2 Diabetes Genetic Loci Informed by Multi-Trait Associations Point to Disease Mechanisms and Subtypes: A Soft Clustering Analysis." *PLoS Medicine* 15 (9): e1002654. https://doi.org/10.1371/journal.pmed.1002654.

Xiang, Zhao-Yu, Shu-Li Chen, Xin-Ran Qin, Sen-Lin Lin, Yi Xu, Li-Na Lu, and Hai-Dong Zou. 2023. "Changes and Related Factors of Blood CCN1 Levels in Diabetic Patients." *Frontiers in Endocrinology* 14 (June): 1131993. https://doi.org/10.3389/fendo.2023.1131993.

Yu G, Wang L, Han Y, He Q (2012). "clusterProfiler: an R package for comparing biological themes among gene clusters." *OMICS: A Journal of Integrative Biology*, **16**(5), 284-287. doi:10.1089/omi.2011.0118.