# CGS4144 - Team 12

## Abstract

## Type 2 Diabetes (T2D) is the most common form of diabetes, wherein the pancreas does not make enough insulin, the body does not use insulin properly, or both. Within the pancreas, islet cells are typically responsible for producing hormones that regulate bodily functions and maintain normal blood glucose levels. Malfunction of these pancreatic islet cells typically underlies the development and progression of T2D. Previous research has explored the relationship between T2D and gene expression, though as diabetes remains uncured, any additional insight into the pathology of the disease and the relationship of specific genes could help clarify previously unseen connections (Xin et al 2016). Using the same dataset, we investigate the question: What genes are pertinent to type two diabetes in pancreatic islet cells? Here we show that T2D is tied to changes in expression of a few key genes across our samples. We are able to identify these key genes using supervised statistical modeling, which opens the door to further research focusing on identifying the precise role and mechanisms these genes and their associated proteins have in T2D. One of the more recent debates in T2D is its broad systemic nature VS specific connection to the pancreas. By identifying key genes in the pancreas, we expand this front of knowledge and potentially even introduce the possibility of more personalized medical intervention by providing a set of significant genes to be considered when designing such treatments.

## Introduction

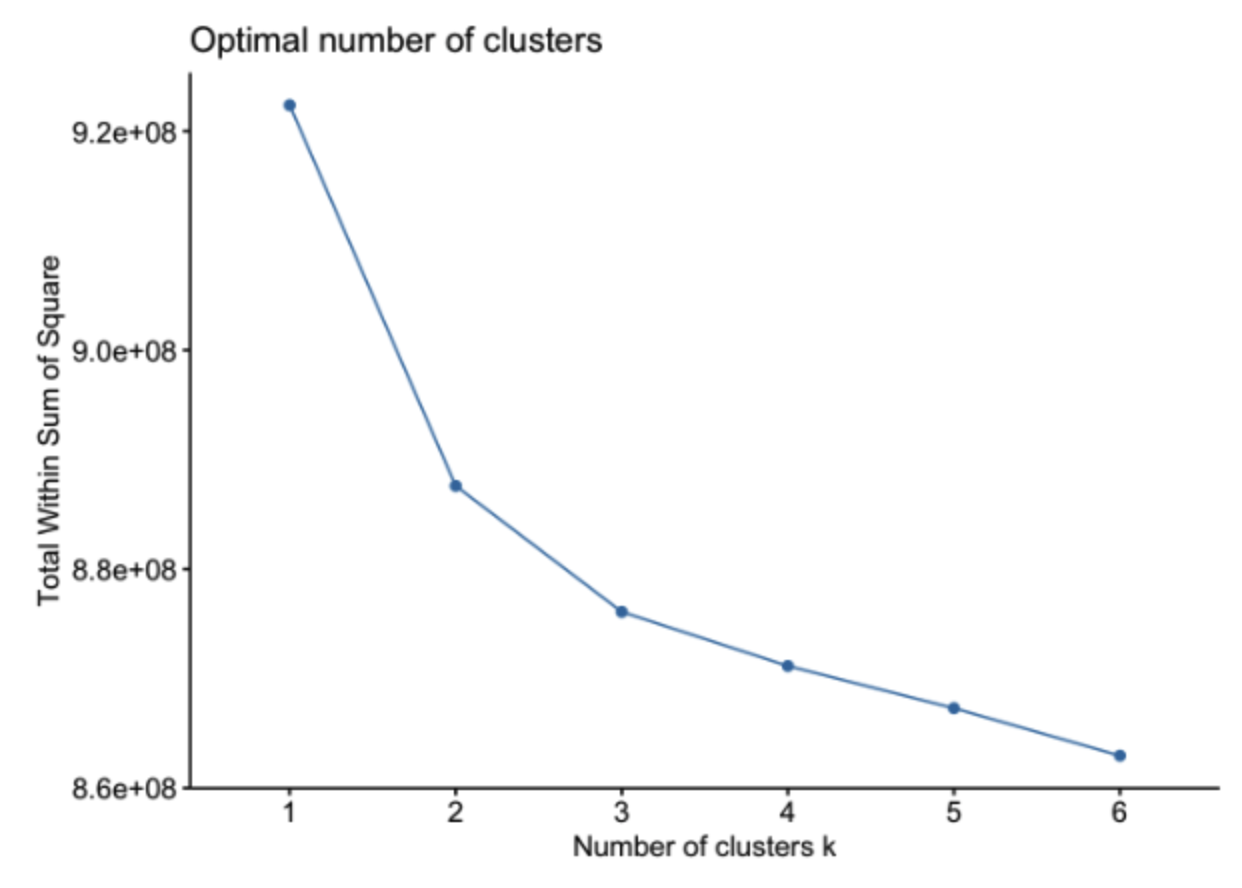
## We utilize a publicly available dataset from refine.bio (Xin et al 2016), consisting of gene expression data from 1600 pancreatic islet cell samples from diabetic and non-diabetic humans. Initially, we sought to investigate the question: What genes are pertinent to type two diabetes in pancreatic islet cells, and how are they related across humans and mice? Additionally, we hoped to investigate and validate the results presented in the dataset’s affiliated paper. Unfortunately, the provided dataset only contained data from human samples, constraining our investigation to genes that help identify the presence of T2D in human pancreatic cells only. To conduct our investigation, we ran enrichment analysis techniques and conducted unsupervised and supervised analysis through various clustering and predictive modeling methods.

## 

## Methods

1. **gProfiler: GOst**  
    The functional enrichment analysis was performed using gProfiler’s g:GOst tool, an online tool that attempts to display genes clustered by various ontologies, ontologies being such significance as the molecular function to which they contribute, regulatory effects they perform, or relationship with a disease among other categories (Kolberg et al 2023). After isolating statistically significantly differentially expressed genes in R, we pasted the data into g:GOst, using the resulting output to explore three different ontologies: Biological pathways (GO:BP), Molecular Function (GO:MF), and Cellular Components (GO:CC). We then analyzed the resulting table of enriched processes along with the p-value provided for each process within each ontology, discussing possible relationships and pointing to related papers that defend our educated guesses as to these. We additionally use GOst in assignment 4 to explore the top discriminating genes that SVM was able to identify (see [Results](#_660hp6wqyisd)).  
     
   [See figures A6-A6.3 in the Appendix for further results](#_660hp6wqyisd) from Assignment 1 GSEA.
2. **K-means Clustering**  
    K-means clustering is a method of clustering which looks for a fixed number (K) of clusters within a provided dataset. To begin, the algorithm creates randomly assigned centroids, which in turn determines the position of the clusters. For each sample, the distance to every cluster’s center (also known as the centroid) is calculated, and the sample is assigned to the closest existing cluster centroid. The algorithm then optimizes these clusters, recalculating the centroids and iterating until the centroids have stabilized or the maximum number of iterations has been reached.   
    To determine the ideal number of clusters (K) for our data, we first utilized the fviz\_nbclust command provided by factoextra. Given a dataset and a maximum number of clusters, this command will calculate K-means clustering with different K values as well as the total within the sum of squares for each clustering attempt. The total within the sum of squares explains proximity of points to their centroid. Lower values mean tighter groupings, indicating a better “goodness of fit”.   
    We then use the elbow method (Figure 1) to determine the optimal number of clusters (K) from the resulting graph. The elbow method means we look for where the elbow in the curve occurs, picking a number of clusters K that does not overfit the data, but also provides a significant reduction in WSS compared to K-1 clusters. Here we determined 3 to be the optimal K value. Afterwards, we used the “kmeans” command to run K-means clustering on our data, setting K to 3.

Figure 1: Elbow Method, K-Means, WSS



Results: K-Means does not directly contribute to the results outlined in this report as the clusters it identifies, per Chi Square testing (See Figure A11), are not aligned with the classes we are attempting to identify, as an unsupervised method tends to pursue stronger trends that differentiate samples than the ones associated with our classes of diabetic/nondiabetic individuals. Regardless, when feature vectors were sufficiently extended to include less variable genes (N=10,000), we observed a low WSS of 4.2%, showing KMeans works as an unsupervised method, just not for our purposes.

1. **Hierarchical Clustering**

Hierarchical clustering is a method in data analysis and statistics that organizes objects into a tree-like structure based on their similarities. The algorithm begins by treating each data point as an individual cluster and progressively merges the most similar clusters until all items belong to a single cluster at the root. This creates a hierarchy or dendrogram, where the proximity of branches indicates the degree of similarity between clusters.

In order to determine the optimal number of (K) clusters for our dataset, different k values were tested. K values ranging from 2-4 were tested and the value 2 was determined to be most suitable because our dataset contains two main divisions in our diabetes testing samples. The fviz\_cluster function was used to visualize clustering results (See Appendix: Figure A8.1). When the K value was tested with values greater than 2, overlap was not significantly reduced.

Results: HCluster does not directly contribute to the results outlined in this report as the clusters it identifies, per Chi Square testing (See Figure A11), are not aligned with the classes we are attempting to identify and produce P values approaching 1.0.

1. **PAM Clustering**

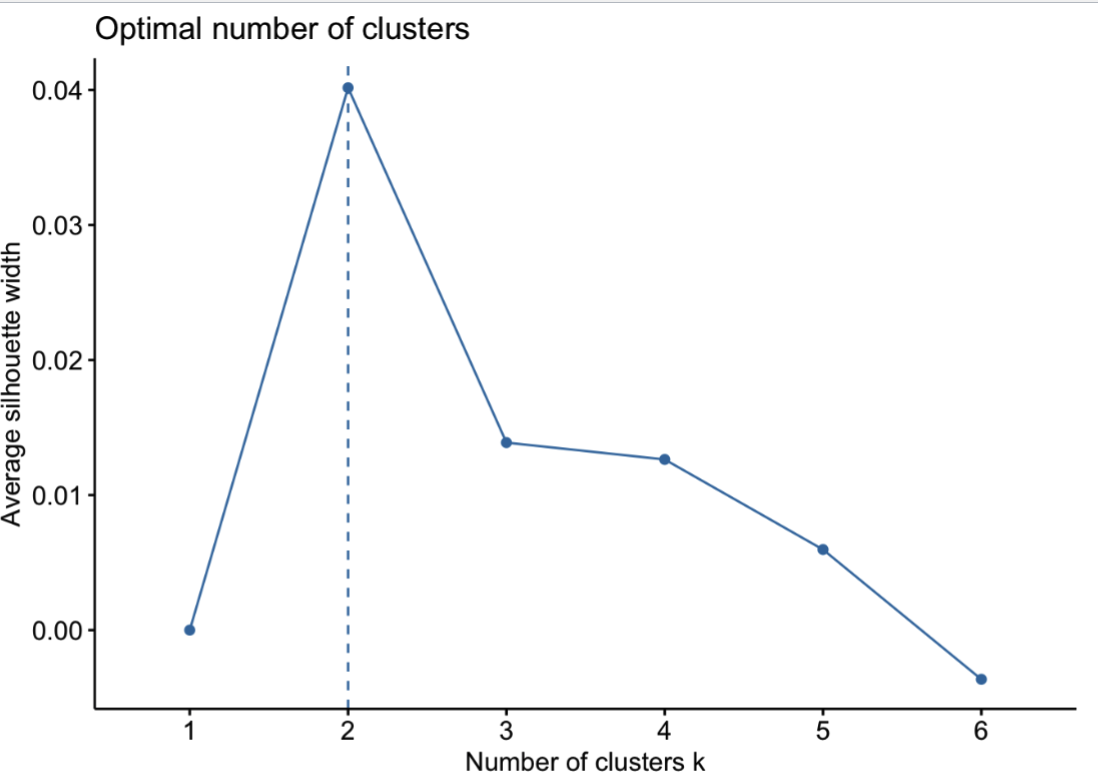
The Partition Around Medoids (PAM) algorithm is considered a more robust version of the K-means algorithm. Similar to K-means, PAM clustering searches for a set number (K) of clusters in a given dataset, searching for the optimal distance between data points and a cluster’s center. PAM clustering sets actual data points as the center of a cluster, known as medoids. This is in contrast to K-means clustering, where the centroid of a cluster is not necessarily an actual data point. This property makes PAM clustering much more robust against outliers in comparison to K-means clustering.

To implement PAM Clustering, we feed in the most significant genes to the pam function as data, along with the number of clusters, K. We performed clustering on all sets of data with K set to 3, and compared PAM clustering on the top 5,000 most significant genes at K values of 3, 4, 5, and 6 (See Figures A9.1-9.4)

Similar to K-means, we also determined the ideal number of clusters, this time using the silhouette method with the fviz\_nbcluster function. For this, we fed in the top 5,000 most significant genes as the dataset, “pam” as the method parameter, and “silhouette” as the method. The resulting graph is shown below.

With the silhouette method, we can validate our clusters by measuring how well an observation is clustered and estimate the average distance between clusters. The higher the value, the more well clustered a given data set is. However, as these values are on a scale of 0 to 1, it is clear that even the average silhouette width at a K value of 2 is extremely low.

Figure 2: Silhouette Method, PAM



Results: PAM does not directly contribute to the results outlined in this report as the clusters it identifies, per Chi Square testing (See Figure A11), are not aligned with the classes we are attempting to identify and produce P values approaching 1.0.

1. **Support Vector Machine**

Support Vector Machines (SVM) are a supervised learning method typically used for classification and regression analysis. To implement SVM for predictive modeling of our data in R, we created a task from the mlr3 package, providing a subset of our data as the “backend” parameter, the “diabetes” column as the target column, and “Diabetes” as the positive value (Lang 2019). Our data was then split (70% for training, 30% for testing), and used to train a model using a Learner. The Leaner was set to be a SVM classification model (“classif.svm”), with the kernel set to linear (“linear”) and the prediction type set to probability (“prob”). Afterwards, the model’s predictive capabilities were tested on the test set. Several hyperparameters exist for SVM but defaults were used for our purposes. Additional kernels were tested but performed worse and were not included in the results. A linear kernel is the only adjustment we make to the default radial kernel. Otherwise, a Nu of 0.5 is used, epsilon of 0.1, tolerance of 0.001, and cost of 1 is used. Further documentation of hyperparameters and defaults can be found at [mlr3learners.](https://mlr3learners.mlr-org.com/reference/mlr_learners_classif.svm.html) For results, please see results below, as SVM was significant and can be used to draw conclusions relating to the goal of this report, unlike unsupervised methods or less performant supervised methods explored below.

1. **Logistic Regression**

Logistic Regression is a discriminative supervised model, often used for classification tasks. To conduct logistic regression in R, we utilized the mlr3 package (Lang 2019). We created a task using the exact same parameters as those used for the SVM task, for the most relevant parameters. Following this, we split our data (70% for training, 30% for testing) and trained our model using a Learner set to be a logistic regression model for classification ("classif.log\_reg") with probability (“prob”) set as the prediction type so that coefficients could be later extracted. Once the model finished training, the model was used on the test set, feeding the task and test data into the Learner’s “predict” function.

For results, see figure A12.2. Logistic regression suffers from overfitting and dramatically loses performance as feature vectors are expanded in size to 1000, 5000, and 10,000 genes. This is partly due to the collinearity of various expressed genes in the data but is also a consequence of our choice to train strictly on most variable genes, which means our data is, by design, that of the noisiest features. We hope this will produce better predictions, but in reality the model learns this noise when exposed to sufficiently high-dimensional feature vectors, and thus fails spectacularly at classification on the test set.

1. **Random Forest**

A Random Forest is an ensemble learning technique used in machine learning for both classification and regression tasks. It constructs a multitude of decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. Each tree is built using a random subset of the training data and a random subset of features, introducing diversity and reducing overfitting. The randomness in data selection and feature choice allows the Random Forest to capture complex relationships within the data and make robust predictions. The final prediction is a result of aggregating the predictions of all individual trees.

The diabetes column of the dataset was used as the target column. 70% of the data was used for training and 30% of the data was used for testing. The model was trained using the mlr3 package in R (Lang 2019). The specific implementation of the Random Forest algorithm for classification tasks that the mlr3 package uses is “ranger” (Wright 2017).

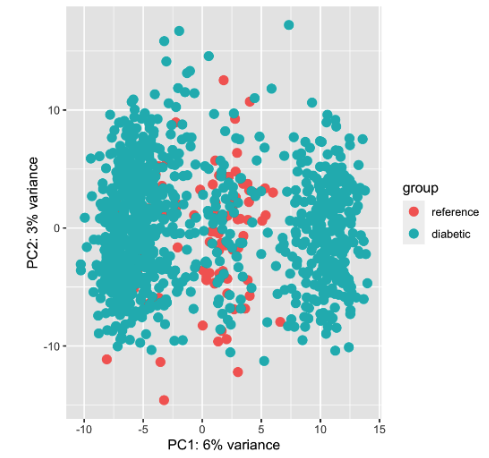
For results, see Figure A12.3. The best AUC score obtained was at feature vectors of the top 100 most variable genes, not unlike logistic regression which is also sensitive to overfitting, after which performance declines as a consequence of overfitting and the “curse of dimensionality” to which random forest is exceptionally prone to. Tweaking hyperparameters to expand cluster size or interrupt the algorithm at lower tree heights may produce better results, but these hyperparameters were not explored as SVM is known to be significantly more performant across the board in classifying within high-dimensional data.

We note that all of the code used for the above methods and our analysis is available on our Github repository, located at [this link](https://github.com/Skunkmeister/CGS4144-Team12).

## Results

Though we explored various supervised, unsupervised, and enrichment analysis statistics, we must discard our unsupervised results as these identify trends that are generally misaligned from the supervised clustering methods for which the trends describing T2D-related gene expression are far weaker and go ignored in unsupervised analyses. This is especially evident in the PCA chart below showing how there are clearly large trends in our genes (PC1 Axis) that do not effectively capture variation across labeled groups.

Figure 3: PCA on Top Variable Genes



We can further defend the claim that unsupervised analysis was fruitless through a Chi-Square analysis of independence of the clusters identified to the labeled groups. What we consistently find is very high P values approaching 1, indicating independence of the groups and no relationship of unsupervised clusters to the reference groups (see Figure A11 for details).

Additionally, enrichment analysis techniques emphasize differential expression without attempting to discover trends or relationships across groups. This means that they only yield exceptionally expressed genes, which alone do not provide meaningful information beyond the role of the pancreatic islet cells should their expression be compared to that of other groups. Thus, for our results section, we turn primarily to the findings of our supervised approaches. In order to remain compliant with the “Results: Method” section of the rubric, we have outlined insignificant results and intermediate observations in the “methods” section.

Of the supervised approaches, the most performant in terms of Area-Under-Curve (AUC) was SVM when performed on top 10,000 most variable genes, with an AUC of 93.62%. This is unsurprising given SVM is known to perform well in high-dimensional data. Using it we were able to achieve a true-positive rate of 95.4% and a false-positive rate of 54.8%. Using SVM we are able to extract genes that contribute the most to the learned prediction. From the 10,000 genes used, we isolate the top 22 (arbitrarily, the cube root of 10,000). These are:

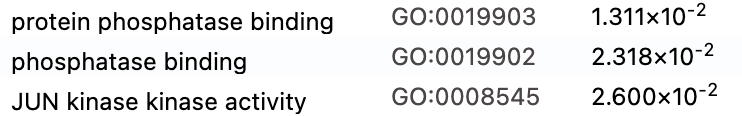
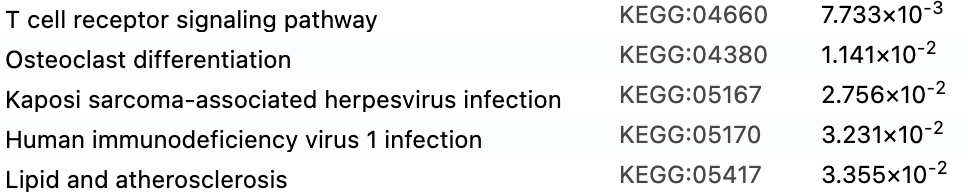
ENSG00000076984, ENSG00000107758, ENSG00000185728, ENSG00000234432, ENSG00000161714, ENSG00000255468, ENSG00000235888, ENSG00000117528, ENSG00000146859, ENSG00000143842, ENSG00000125107, ENSG00000151729, ENSG00000016864, ENSG00000245573, ENSG00000110442, ENSG00000104154, ENSG00000081307, ENSG00000271590, ENSG00000180758, ENSG00000104714, ENSG00000267321, ENSG00000108344

We have therefore achieved our goal of finding genes the expression discriminates and therefore correlated with our label of diabetic/nondiabetic individuals by training an SVM model on 70% of samples with feature vectors consisting of the top 10,000 genes.

While we are confident that these genes are quality discriminators, we note the intriguing lack of overlap between top discriminating genes as we truncate the feature vectors fed to SVM. However, we anticipate this is primarily a consequence of the colinearity of various genes’ expression. As we introduce more of the top variable genes, we introduce genes that while being weakly expressed compared to those at the top of the feature vector, may still play a more significant role in diabetes from even comparably insignificant levels of expression or more consistent expression across samples. Therefore, we discard this as a significant concern.

In our process we made no further significant discoveries in the data, though we were able to run enrichment analysis on the genes above, identifying their connection primarily to various signaling pathways and, evidently, Alzheimer disease. When we perform an ordered query that utilizes SVM’s ordering of the weights, we are able to additionally identify association with atherosclerosis, HIV infection, Herpes, and a handful of basic biochemical pathways like phosphatase binding.

Figure 4: GSEA on Top SVM Discriminating Genes



Please see appendix for further, less significant results and descriptions as preserved from assignments per instruction.

## Conclusion

Going into this project, we hopefully hypothesized that we may be able to identify genes which contribute to / are correlated with T2D. We are proud to have succeeded, though having simultaneously identified that unsupervised clustering of variably-expressed genes as well as enrichment analyses without first segregating by label lends little insight into the disease we are hoping to identify.

In assignment two, we identified the functions of our most variably expressed genes through enrichment analysis using GOST. However, we drew no insights towards their relationship to diabetes, only assuming that variable expression might be correlated. We did later find this to not be the case, as evidenced by the coefficient instability in SVM as feature vectors are expanded to include less variable genes).

In assignment three, we learned that unsupervised clustering, though very effective at identifying large trends in data and segregating our genes, fails to identify the trends explaining our independent variable of diabetes, as evidenced by extensive chi square testing.

Finally, in assignment four, we additionally discovered that some other supervised methods like logistic regression and random forest are prone to overfitting and struggle with high dimensional data compared to SVM.

If we were to pick an optimal supervised learning method, rather than exploring several different ones, we would attempt to tune the hyperparameters of SVM after identifying the most significant genes with PCA. We could then use euclidean distance to reintroduce collinear, but therefore related, genes to our "signatures" and visualize these exclusively. As random forest and logistic regression both perform very poorly in comparison, this would likely result in significant improvement.

Additional analysis would be wise. In the future we feel applying an MLP / Deep Learning to this classification task, though computationally intensive, might allow us to better predict diabetes from the dataset or at the very least approach a theoretical maximum accuracy achievable from the dataset. It also opens the door to ablation studies which have the possibility of reducing overfitting without sacrificing accuracy in a rather controlled manner (potentially less sensitive or more granular than hyperparameter tweaking in an SVM).

## References

Lang M, Binder M, Richter J, Schratz P, Pfisterer F, Coors S, Au Q, Casalicchio G, Kotthoff L, Bischl B (2019). “mlr3: A modern object-oriented machine learning framework in R.” *Journal of Open Source Software*. [doi:10.21105/joss.01903](https://doi.org/10.21105/joss.01903), <https://joss.theoj.org/papers/10.21105/joss.01903>.

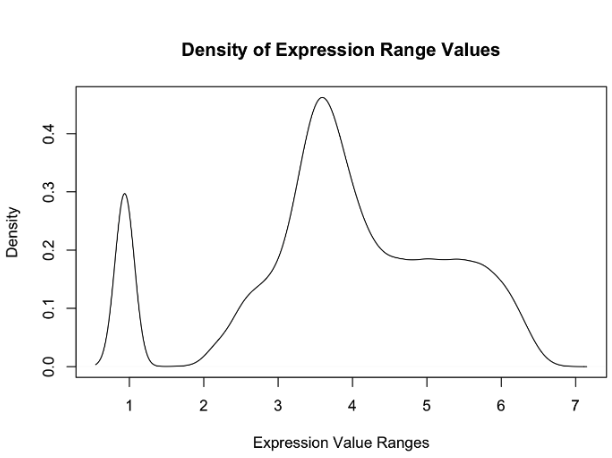
Kolberg, L., Raudvere, U., Kuzmin, I., Adler, P., Vilo, J., & Peterson, H. (2023). G:Profiler—interoperable web service for Functional Enrichment Analysis and gene identifier mapping (2023 update). *Nucleic Acids Research*, *51*(W1). https://doi.org/10.1093/nar/gkad347

Wright, M. N. & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. J Stat Softw 77:1-17. tools:::Rd\_expr\_doi("10.18637/jss.v077.i01").

Xin, Y., Kim, J., Okamoto, H., Ni, M., Wei, Y., Adler, C., Murphy, A. J., Yancopoulos, G. D., Lin, C., & Gromada, J. (2016). RNA sequencing of single human islet cells reveals type 2 diabetes genes. *Cell Metabolism*, *24*(4), 608–615. https://doi.org/10.1016/j.cmet.2016.08.018

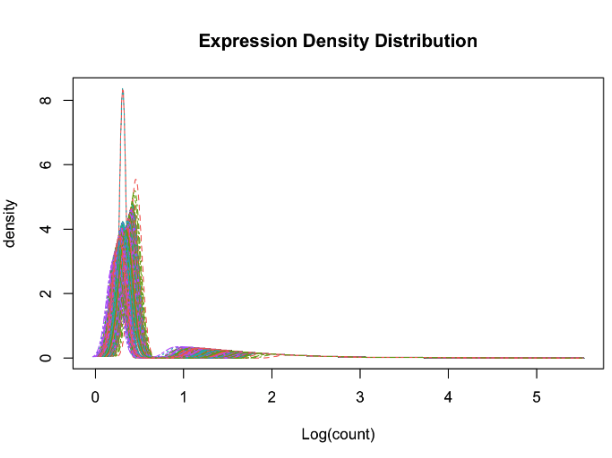
## Appendix

### Figure A1:



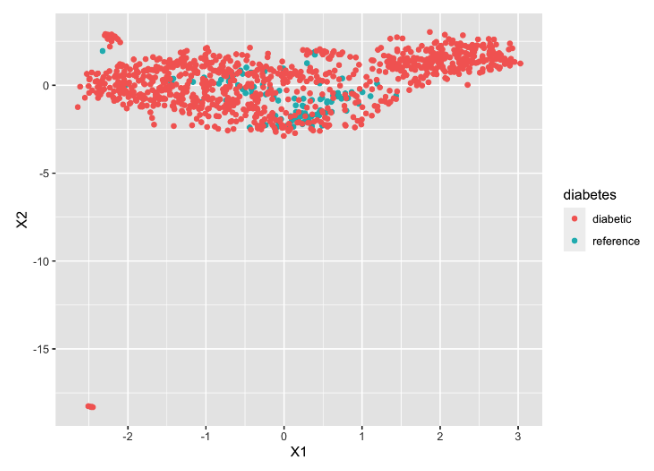
The range density plot shows an intriguing bimodal distribution. A large group of genes has a minimal range, whereas the majority of the genes fall close to a range of 3.5 (log scale) in expression between samples. Genes with a low variability across all samples are not interesting since this data set contains both diabetic and control patients.

### Figure A2



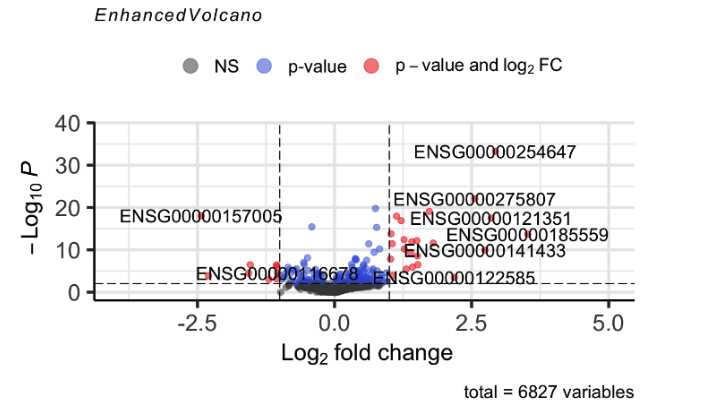
The majority of the genes show negative median log-scaled expression (0 to 1), simply implying minimal expression relative to other genes. However, a group can be seen with values near 1-2. These imply that the genes in this group are those which, relative to other genes gathered from the pancreatic islet cells samples, are being expressed the most. This stands against no particular control, however, and so does not yield any meaningful conclusions of the genes at hand being diabetic, but rather just genes that are expressed more.

### Figure A3: UMAP



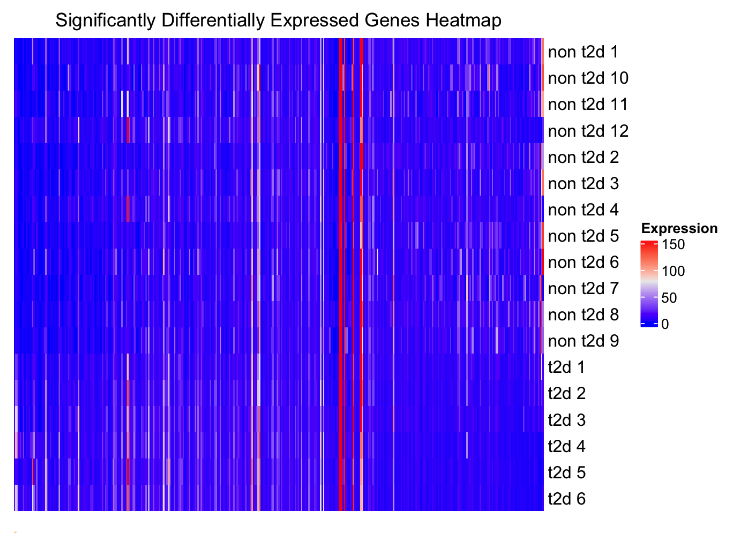
The UMAP plot shows that the two groups are clustered together in distinct groups. These clusters are the two different cell types, diabetic and reference. There are a few outliers in the plot where a few samples are displayed at the bottom of the plot. UMAP is a nonlinear technique that excels at preserving local structure, making it well-suited for exploring complex relationships and identifying clusters in high-dimensional data.

### Figure A4: Volcano Plot



It seems a handful of genes have stood out as having significant fold change in both the positive and negative direction relative to the control. Of the 6827 genes examined, 357 meet the threshold discovered for significant change in differential expression, visualized as north of the gray line towards the bottom of the volcano plot. Some are exceptionally negative / positive, and are labeled for their uniqueness.`

### Figure A5



Heatmap created using ComplexHeatmap showing the differentially expressed genes. Side bar added that shows sample groups (cancer vs not).

To create this Heatmap, we used ComplexHeatmap and simply aligned our raw data with the significantly expressed gene table, then fed in only the significantly expressed gene data to ComplexHeatmap. We also grouped this data by the subject, with “non t2d” denoting a subject without Type 2 Diabetes and “t2d” denoting a subject with diabetes. We found that ComplexHeatmap’s default settings clustered the rows and columns so that they appeared in a different order than how they appear in the input table. To fix this, we set cluster to FALSE for both rows and columns.

Based on this heatmap from our significantly expressed gene data, there does appear to be some degree of difference between how certain groups of genes are expressed across the subjects with diabetes and those without. Interestingly, there appear to be differences in how diabetes is expressed within the diabetes groups, with subjects 1,2, and 3 differing from subjects 4,5, and 6. A similar division within the non-diabetes group, if it exists, does not appear to be quite as apparent.

### Figure A6: Gene Set Enrichment Analysis

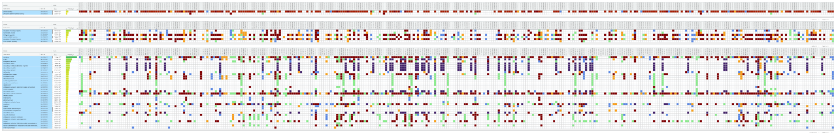
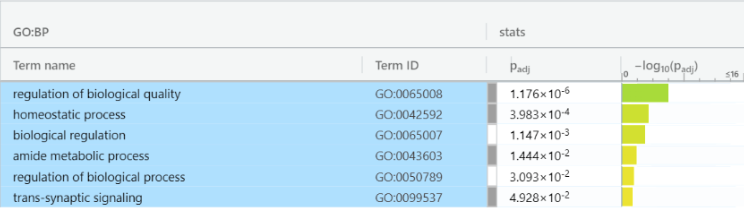
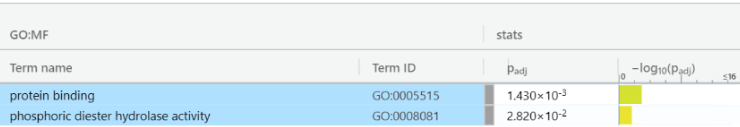
Figure A6.1 (All) 

Figure A6.1: GSEA: Biological Pathways



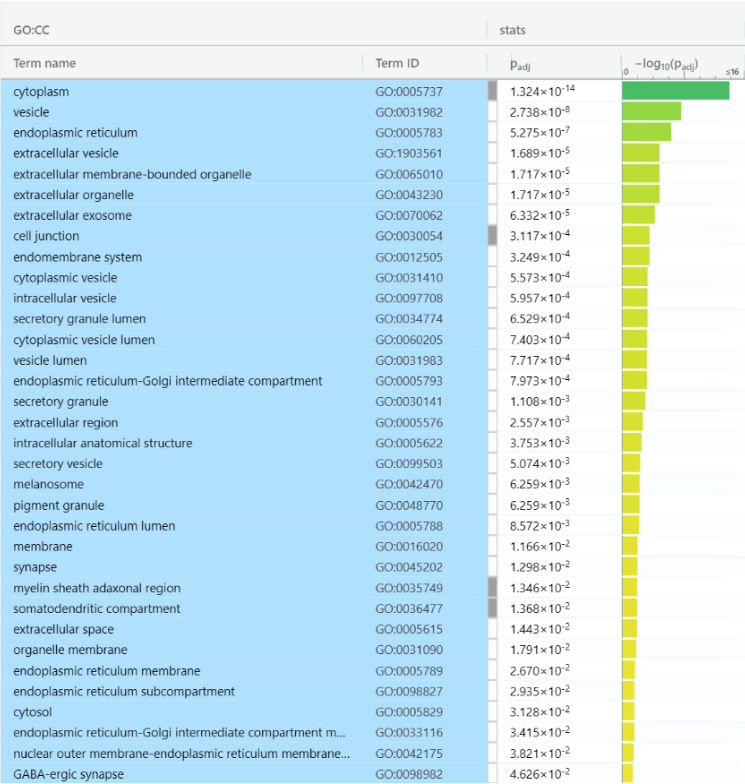
Biological quality regulation was also closely tied to the differential expression of our genes, under the biological pathway ontology. This too is somewhat unsurprising, given this disease might be tied to underperforming pancreatic islet cells. Particularly the relation to “amide metabolic process” is more specific than the other points noted in the table, and prompted some investigation. This paper (https://pubmed.ncbi.nlm.nih.gov/9421375/) seems to connect amides (which we recognize are just a molecular component) to insulin secretion, which confirms some of the findings from the cellular component ontological exploration above.

Figure A6.2: GSEA: Molecular Functions



Finally we see a connection to protein binding and phosphoric diester hydrolase activity within the molecular function ontology. This is intriguing as it may relate to insulin receptors’ capacity to bind free glucose, though it is our understanding that this is a fundamental quality not only of pancreatic islet cells but of all somatic cells participating in metabolism. Instead we explore phosphoric diester hydrolase activity in an effort to supplement our understanding: It turns out that phosphate diester hydrolases are closely tied to cAMP regulation which we know from AP Biology regulates the cell cycle in some capacity. When searching specifically for a connection to pancreatic islets however, we find [the following paper](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5587329/#:~:text=In%20addition%20to,conjugation%20with%20glutathione.) which seems to suggest a connection to type 1 diabetes in which this very class of enzyme’s downregulation contributes to oxidative stress and general dysfunction.

Figure A6.3: GSEA: Cellular Components



Of most significance is the cytoplasmic relationship within the GO:CC (GO:CC) table. We can see a very significant P value here of 1.324x10^-14, which is seemingly reasonable since diabetes would influence certain metabolic pathways naturally present in the cytoplasm, or so we would assume from the results. This led to a further exploration on google scholar, which in turn led to [this paper.](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8115730/), suggesting that NADPH is depleted and lipogenesis from glucose dampened as a consequence of T2D in pancreatic islet cells. We can also see that functions relating to vesicles are affected, which falls in line with the paper’s suggestion that secretion is hampered (note the secretory vesicle , around halfway down in terms of significance).

### Figure A7: K-means Cluster Plot

The cluster plot shows how these 3 clusters look when displayed in a 2D graph via PCA. The purpose of this is that it serves as a check to make sure that using four clusters is an acceptable decision. The graph does appear to show overlap between these three clusters.

### 

### Figure A8 (Hierarchical Clustering)

Figure A8.1: 2 Clusters

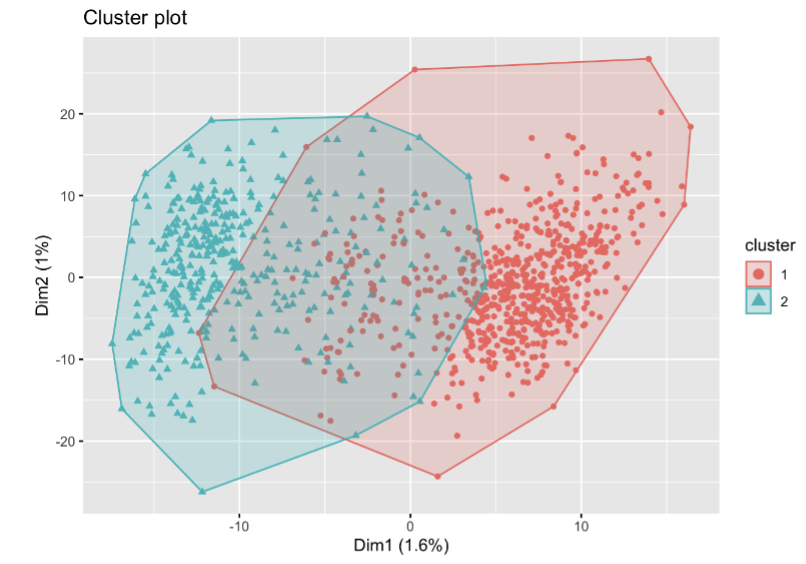
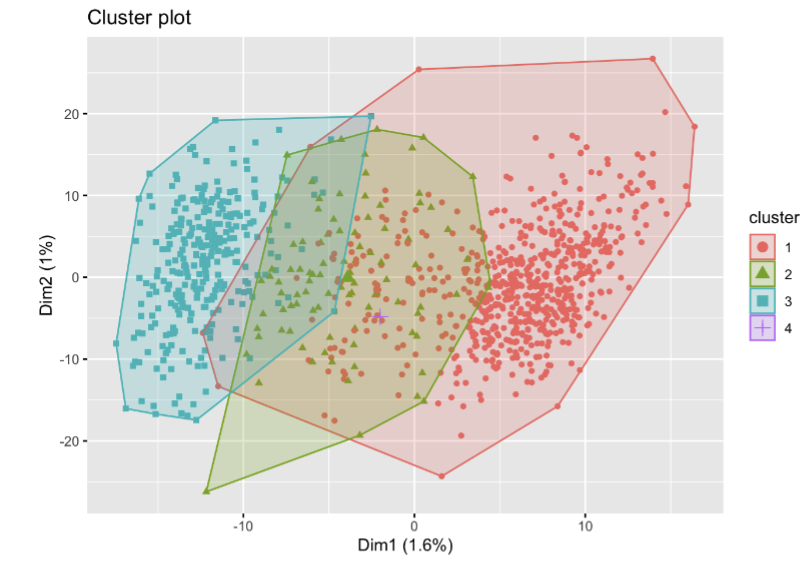


Figure A8.2: 3 Clusters

Figure A8.3: 4 Clusters



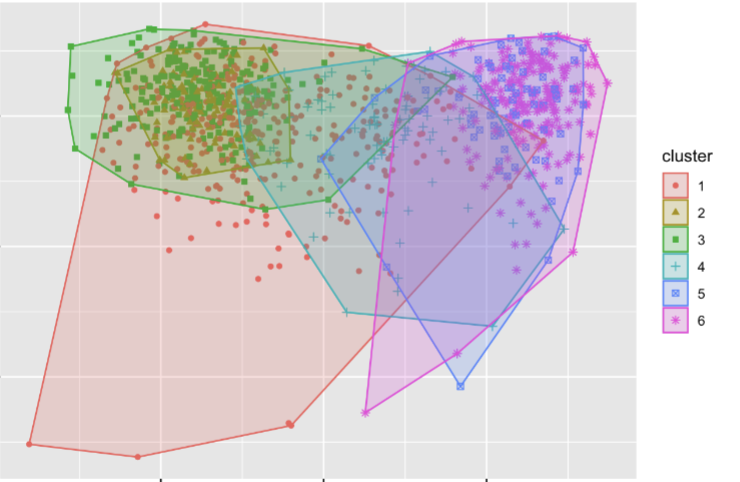
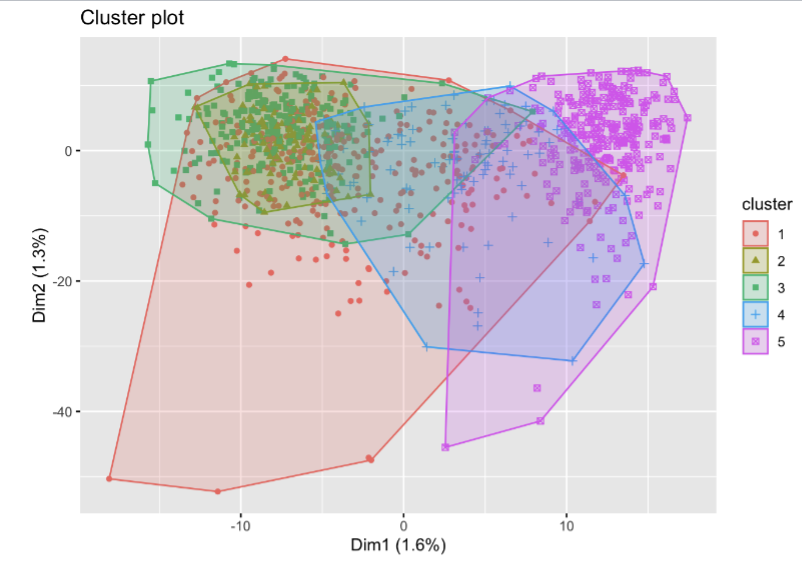
This set of graphs is quite interesting. The two cluster graph amusingly displays what appears to be a single cluster, likely another folly of the 2D world. The three cluster graph shows how the larger visible cluster has bifurcated (this time quite literally given the nature of hierarchical clustering). The four cluster graph further divides these two larger clusters, though it is unclear from which the points of cluster 1 originate, but the extremely tight cluster (now cluster 4) remains centered and nearly invisible.

### 

### 

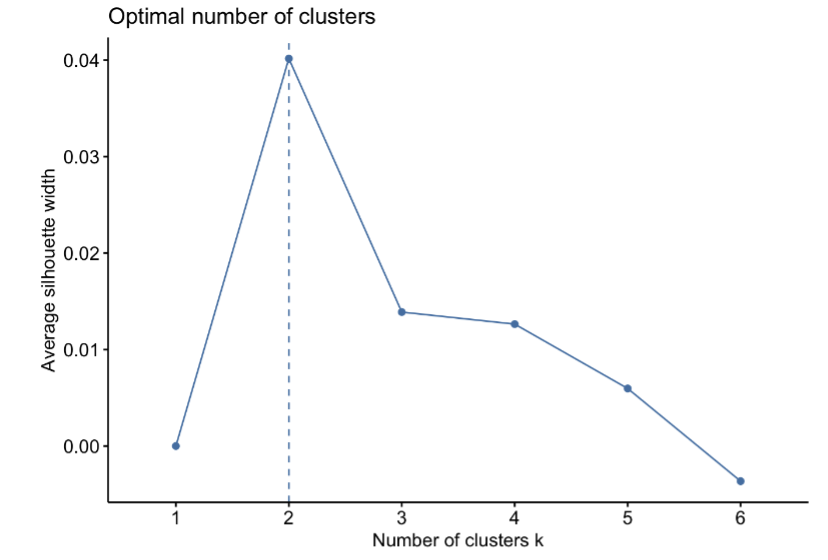
### Figure A9 (PAM Clustering) (9.1 - 9.4)

### 



We can see how membership changes above, at least in the two chosen dimensions for the visualization (since we’re operating on 5000-dimensional data). Indeed it seems that more clusters reduce the overlap, but the extent to which they reduce the overlap is better demonstrated through the silhouette method shown below, because, amusingly, the new clusters introduced in the 2D field above show literally 0 reduction in overlap. 4 of the generated clusters overlap almost perfectly with each other, implying the PCA has failed to adequately visualize these groups.

Figure A9.5: PAM Clustering Goodness of Fit via Silhouette Method

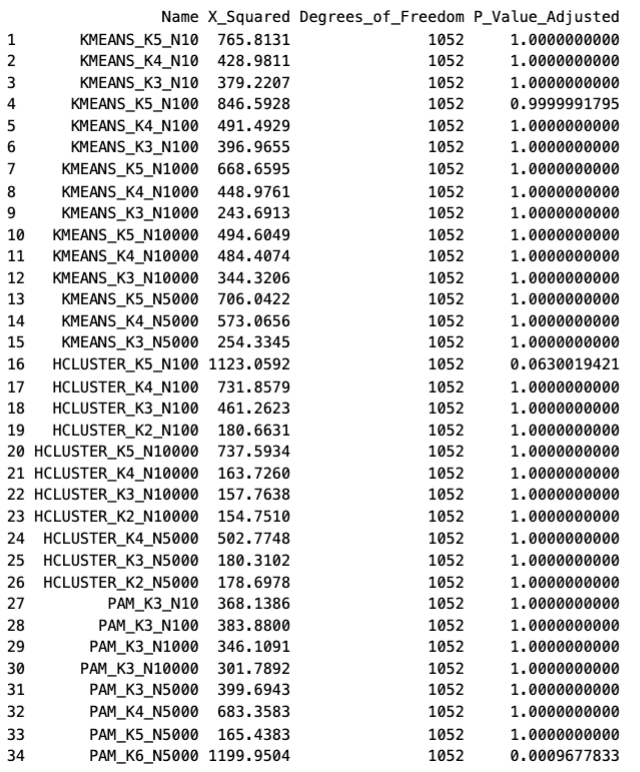


The silhouette method used above is looking at all the points resulting from different K values of PAM, giving them a score (based on their proximity to their assigned cluster and distance to other clusters), and then averaging the score across all points to determine a total score that represents discrimination. We’d expect this to increase with more clusters but alas it seems 2 clusters are locally optimal in effectively segregating our samples.

### Figure A10: Unsupervised Clustering Heatmap

To generate the above heatmap, we utilized the ComplexHeatmap library again using the top 5,000 genes as instructed. We provided a legend based on gene expression, and to cluster the genes within the heatmap, we set the cluster\_columns and cluster\_rows parameters to “TRUE”. This also displays the dendrograms for the rows and columns. Unfortunately, we were unable to cluster the heatmap according to the clusters we determined in earlier sections of the assignment, and so were unable to annotate sections for said columns. While cluster data do clearly exist, the heatmap unfortunately does not clearly reflect this fact.

### Figure A11: Chi-Square Tests of Significance



It is safe to assume that the groups chosen are indeed independent from the reference group of diabetic / nondiabetic individuals with rare exceptions in K5 Hierarchical clustering of top-100-most-variable-gene feature vectors and K6 PAM of the top 5000. This may have to do with the fact that we chose the most variably expressed genes, not the most expressed genes period, which generally picks the genes noisiest across our samples and thus yields relatively poor clusters (as seen by the oddly low amount of clusters chosen by silhouette method or the lack of an elbow in the WSS K-means evaluation)

### 

### Figures A12: Predictive Modeling / Supervised Methods

Figure A12.1: SVM

| Top N Variable Genes Feature Vector Length | TPR | FPR | AUC |
| --- | --- | --- | --- |
| 10 | 100 | 100 | 61.9 |
| 100 | 99.6 | 100 | 83.3 |
| 1000 | 96.1 | 64.5 | 88.9 |
| 5000 | 98.6 | 71 | 89.8 |
| 10000 | 95.4 | 54.8 | 93.6 |

Figure A12.2: Logistic Regression

| Top N Variable Genes Feature Vector Length | TPR | FPR | AUC |
| --- | --- | --- | --- |
| 10 | 97.9% | 90.3% | 84.5% |
| 100 | 92.3% | 35.5% | 88.5% |
| 1000 | 53.0% | 48.4% | 52.3% |
| 5000 | 49.5% | 45.2% | 52.3% |
| 10000 | 56.5% | 45.2% | 56.0% |

Figure A12.3: Random Forest

| Top N Variable Genes Feature Vector Length | TPR | FPR | AUC |
| --- | --- | --- | --- |
| 10 | 98.6% | 100.0% | 82.1% |
| 100 | 100.0% | 100.0% | 85.7% |
| 1000 | 100.0% | 100.0% | 80.7% |
| 5000 | 100.0% | 96.8% | 79.8% |
| 10000 | 100.0% | 100.0% | 78.9% |

### Figures A13 (Supervised Method “Signature Genes” Heatmaps)

Figure A13.1: Unclustered Heatmap

Two heatmaps are presented below. The reason for this is that we have been provided instructions to illustrate hierarchical clustering on both axes but also to provide annotation data. We go to extra effort to order this annotation data in hopes to see contrast on the heatmap that the hierarchical clustering would override the sample order of (because it is susceptible to trends in the provided gene vectors that are unrelated to our target group of diabetics / nondiabetics). In order to ensure we satisfy the conditions of the assignment we have therefore provided two graphs, without and with clustering, to observe these differences. N.B., we do not provide individual column labels as doing so proves impractical. The font of the rows is intentionally small such that they do not intersect making them legible at the rendered resolution.

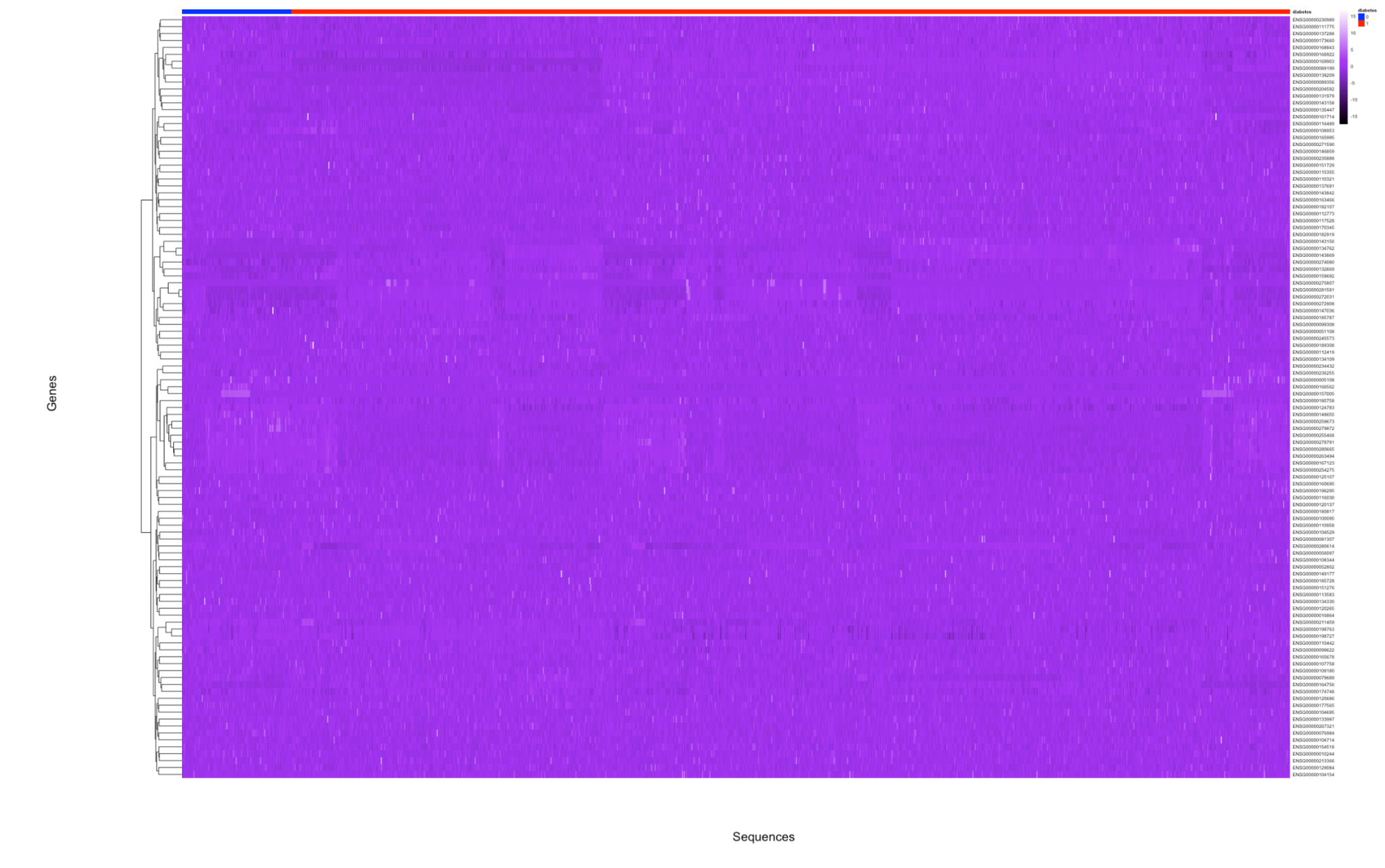
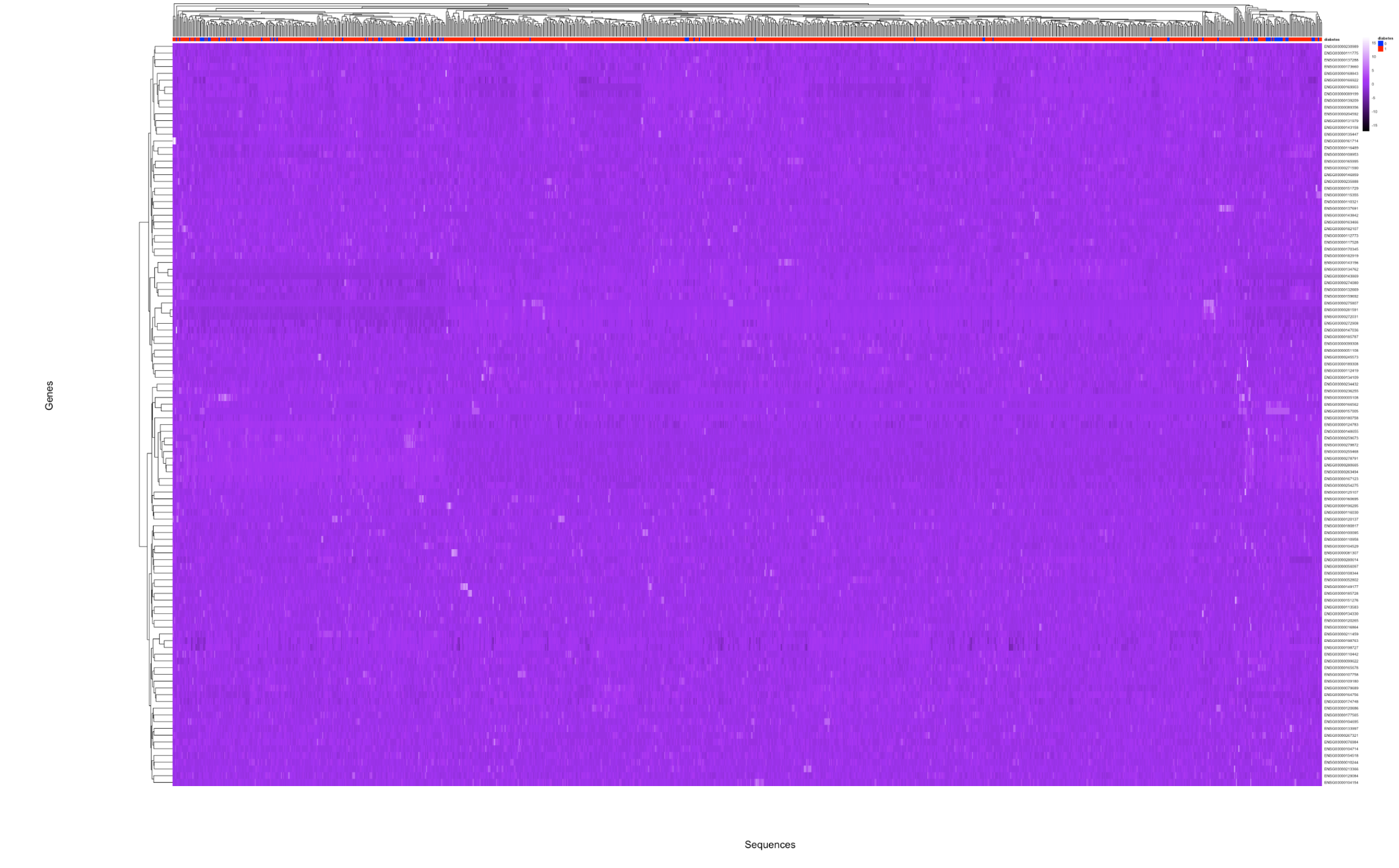


Figure A13.2: Clustered Heatmap



Admittedly the differences in the two heatmaps aren't very pronounced. We scale the data by gene so as to visually present the expression Z score of a gene relative to itself in adjacent samples. Hierarchical clustering does visually present some contrast towards the right of the graph, but our order-by-annotation version of the heatmap (first of the two) is significantly less noteworthy in its trends. We assume that the trends our various clustering algorithms identified were simply more complex than easily made visible. That's to say, SVM likely discovered not individual genes but combinations of genes which help discriminate a sample, and visualizing this relies heavily on the order in which genes are presented. We experimented with row clustering by correlation (cosine / pearson) rather than euclidean distance and otherwise altering the visualization ( scores within samples rather than within genes across samples) but still were unable to show any interesting contrast between our columns of diabetics and nondiabetics in the first graph.