Dimensionality Reduction of Diabetes-Related Gene Expression Data using Variational Autoencoders

Yvon Lu and Icaro Andrade Souza Bacelar

December 14, 2024

Abstract

Understanding the differences in pancreatic islet cell gene expression between individuals with Type 2 Diabetes (T2D) and non-diabetic controls is critical for advancing our knowledge of the disease. In this project, we utilized dimensionality reduction techniques, primarily Variational Autoencoders (VAEs), to analyze single-cell RNA sequencing data from pancreatic islet cells. By constructing an informative latent space, we aimed to identify patterns in gene expression that distinguish diabetic and non-diabetic samples. We observed the impact of including and excluding genes with varying levels of expressivity and investigated the behavior of specific cell types. Despite dataset limitations, our findings reveal the potential of VAEs in uncovering biologically significant latent features.

Introduction

Our study addresses the biological question: How can dimensionality reduction techniques create an informative latent space for analyzing gene expression in pancreatic islet cells? Specifically, we explore whether latent features derived from these techniques can capture the differences between T2D and non-diabetic samples. Prior studies[6] have employed methods like PCA (Principal Component Analysis) and t-SNE (t-distributed Stochastic Neighbor Embedding) for dimensionality reduction, but these models often struggle with capturing nonlinear patterns in high-dimensional biological data.

To overcome these challenges, we leveraged Variational Autoencoders (VAEs), which maps high-dimensional data into a latent space that follows a continuous distribution. VAEs have shown promise in capturing complex relationships in genomic datasets, making them a suitable candidate for this analysis. Our approach builds upon common tools in the field by integrating VAEs with clustering algorithms such as Leiden to evaluate the structure of the latent space.

Methods

Data Collection and Processing

We used dataset GSE81608, which was obtained from GEO's website and contains RNA-seq data of pancreatic islet cells from individuals with T2D and controls. The raw data was presented in a txt format with **1600 columns** (one of each sample) containing close to **40000 rows of gene expression data**. Key preprocessing steps included:

- **Metadata Alignment:** While the dataset did not contain embedded metadata, there was a companion HTML file with information regarding each sample's control/T2D label, age, and gender. However, there was no metadata to allow for the conversion of the gene indexing in the dataset into interpretable gene IDs. By transposing the initial dataset and performing text pairing, we aligned each sample (now rows) with the corresponding metadata.
- **Normalization:** Using Scanpy (as recommended from the previous report) we scaled the data using the function normalize_total, so that every cell had the same total gene count. After tests, we opted to not exclude highly expressed genes when calling normalize_total.
- **Imputation:** Missing values were filled using linear interpolation. We also verified that all samples had above zero total gene count.

- **Balancing:** Used random sample deletion to balance T2D and control samples, leaving **1302** samples.
- **Gene Selection:** Gene expressivity was measured based on variance and total count across samples. Multiple data subsets were formed to train and test models.

Dimensionality Reduction Techniques

- **PCA:** Explore linear patterns by transforming the data into principal components that capture maximum variance. Used as a baseline/floor for metrics.
- **UMAP:** Applied to reduce the high-dimensional gene expression data into a two-dimensional latent space. UMAP was primarily used for visualization and exploratory analysis.
- **VAE:** Implemented using the scVI library, we accessed how modifying parameters led to more informational latent spaces. We utilized Hyperopt for tuning parameters, besides some manual testing. Additionally, we evaluated the use of pertinent modifications, such as the inclusion of Gene Likelihood and Kullback-Leibler Warm-up.

Clustering and Evaluation

- **Leiden Clustering:** Applied to the latent space to identify clusters, ensuring well-connected groups and frequently used in similar research[1].
- **Evaluation Metrics:** Performance was assessed using Silhouette Score, Adjusted Rand Index (ARI), and F-scores obtained with Logistic Regressions, for interpretability, and Support Vector Machine (SVM), to add robustness against non-linerity. Here we used a 80/20 test-train split. Results of papers on similar datasets were used as benchmarks[2][5].

Results

Exploratory Data Analysis

One of the first realizations we came to upon interacting with our dataset was the high degree of variance across gene expressions. The mean values for each of the gene expressions range from as low as 10^{-3} to as high as 10^2, with standard deviations being generally 10 times higher than the mean value for a given gene expression. This raises the hypothesis that a few genes could dominate the expressivity of our data.

By examining genes with the highest variance across samples, we identified key contributors to variability within the dataset. Figures 1 present violin plots showcasing the distribution of expression levels for the 20 most commonly expressed and the 20 most variable genes, respectively. The first plot highlights genes with consistent expression levels across samples, though some exhibit significant outliers, suggesting variability due to biological or technical factors. The second plot focuses on the most variable genes, displaying broader distributions and heavy-tailed behaviors, indicative of extreme values. These variable genes are particularly relevant for clustering and differential expression analysis, as they may drive separation between diabetic and non-diabetic samples. Provide a comprehensive view of gene expression stability and variability, these plots offer valuable insights into potential biomarkers for T2D.

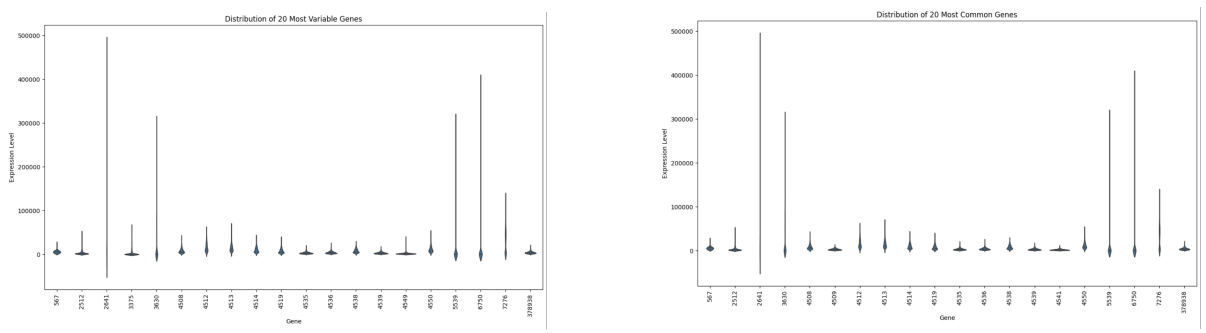

Figure 1: (Left) Distribution of 20 Most Common Genes; (Right) Distribution of 20 Most Variable Genes

Furthermore, by aligning a companion HTML file containing details on each sample with the original dataset and loading it as observation in an AnnData structure, we were able to effectively produce relevant metadata. We noticed the data was unbalanced with roughly 1.5 times the number of diabetic samples when compared to control samples, but double the number of non-diabetic donors to T2D donors, factors that harm the generalization power of our findings. We balanced control/T2D sample count and performed tests in both balanced and original datasets, with balanced datasets leading to marginally better results across the board.

We were also able to link each sample to cell type. The four cell types included are: alpha, beta, delta, and PP. We investigated the proportion of these cells within the data to inform our next steps. As shown in Figures 2, alpha cells dominate, making up 59.1% of our cells. This is followed by beta at 31.4%, and PP and delta at 5.8% and 3.6% respectively. The proportion of these cells for individuals with T2D and our controls is roughly uniform across cell types.

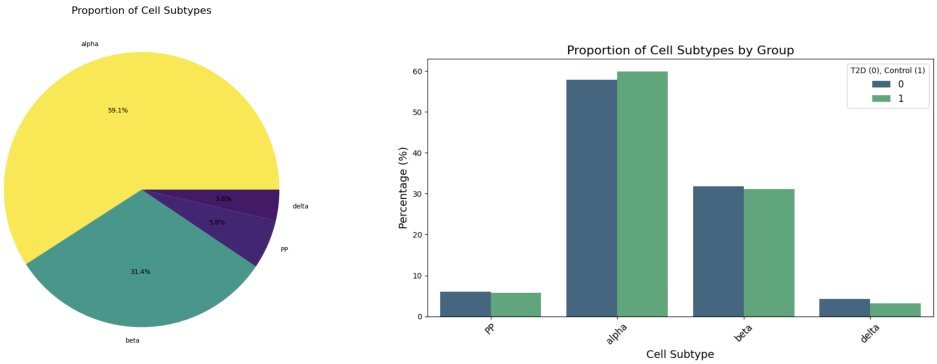

Figure 2: (Left) Proportion of Cell Subtype within Dataset; (Right) Proportion of Group and Cell Subtype within Dataset

We furthered exploration into the data by utilizing a PCA model to identify if simple linear relations of dominant gene expressions could indicate any separation between controls and T2D within our samples. As seen in Figure 3, the PCA graph showed close to no separation indicating that the variance captured by the first two components doesn't adequately distinguish the groups and relevant patterns might be hidden from linear models. We also calculate a Silhouette score to use as a point of comparison for our later analysis. From Xiang et al.[5], a Silhouette score "...measures how well each cell lies with its own cluster, which indicates the separability of each individual cluster. The value of Silhouette coefficient s(i) is between 1 and 1; 1 means that the cell is far away from its neighboring clusters, whereas 1 means that the cell is far away from points of the same cluster...". For this preliminary step our silhouette score is about 0.004, which indicates that PCA is adequate for reducing dimensionality on our dataset. Xiang et al., also tests many dimensionality reduction models on multiple datasets, one with pancreatic islet cells, albeit not using diabetes/control markers. Out of all the datasets they tested, their pancreatic islet cells dataset was the most difficult to work with, with their evaluation metrics being the lowest of the set (PCA Silhouette: -0.017).

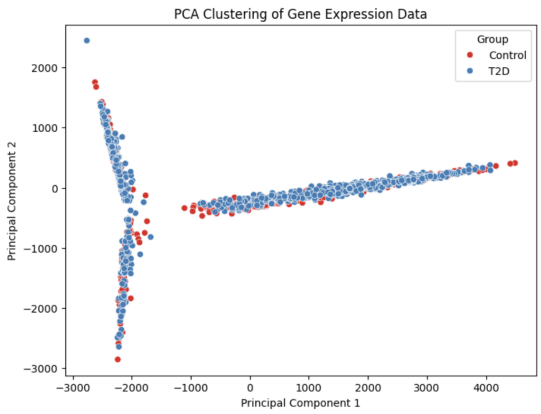

Figure 3: Two-factor PCA Clustering

Upon noticing how a PCA could not effectively capture the patterns in our data, we proceeded to try a similar approach with t-SNE, which is a nonlinear dimensionality reduction technique particularly well-suited for visualizing high-dimensional data. It maps multi-dimensional data to a lower-dimensional space (two in our case) while preserving the local structure of the data. This makes it an effective tool for detecting clusters and patterns in complex datasets. Although we could see clustering, with a Silhouette score of 0.27, as shown in Figure 4, they did not distinguish between control and T2D, with an ARI of 0.001 and F-Score of 0.47.

After seeing how a t-SNE could not cluster our data well, we hypothesized that we needed a model that could better deal with sparse regions in our dataset, addressing sample distance with more flexibility and preserving not only local but also global structure. With that in mind, we tested how our dataset would behave under a Uniform Manifold Approximation and Projection (UMAP) representation, which is another form of nonlinear dimensionality reduction technique designed to capture both local and global data structures. Unlike t-SNE, UMAP excels at preserving the overall geometry of the data, making it particularly effective for uncovering complex patterns and distinguishing clusters, even in sparse or high-dimensional datasets. We observed positive results, with the clusters displaying more separation between diabetic and T2D samples, as seen in Figure 4, with a Silhouette score of 0.31, ARI 0.21 of, and F-score of 0.72. As such, we opted to use UMAPs to visualize our analysis.

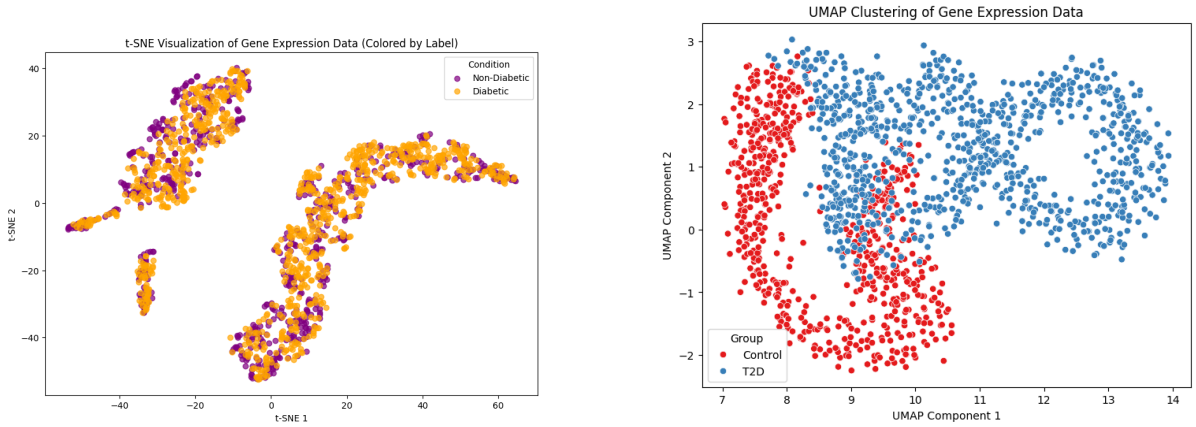

Figure 4: (Left) t-SNE map; (Right) Two-dimensional UMAP

VAE: All Cell Types

We began by implementing a baseline VAE model using the prebuilt scVI library. The initial model employed default parameters and achieved very low scores (Silhouette: <0.03, ARI: <0.03), reflecting its inability to capture meaningful latent features. To improve this, we iteratively enhanced the VAE configuration by varying parameters such as the number of epochs, n_hidden, n_layers, n_latent, and dropout rate. Early stopping with adjustable patience was introduced to prevent overfitting and optimize training time. Hyperparameter tuning was performed using the Hyperopt library, which allowed us to systematically identify optimal parameter settings.

A key finding was the high sensitivity of the Silhouette Score to even small changes in dropout rate, while the number of layers (n_layers) improved scores up to five layers, after which performance declined sharply. We also incorporated Kullback-Leibler (KL) Warmup, which gradually increases the KL divergence term during training. This technique helped prevent the latent space from collapsing early in training, improving the balance between reconstruction loss and regularization. This inclusion led to slightly higher Silhouette score but had no noticeable effect on ARI. We later introduced a zero-inflated negative binomial (ZINB) gene likelihood, which better accounted for the sparse nature of gene expression data by modeling zero counts separately. Furthermore, we replaced K-means clustering with Leiden clustering, as the literature frequently highlights its suitability for single-cell datasets[1].

Despite these improvements, while we managed to obtain Silhouette scores above the threshold of 0.2, ARI scores remained low and F-Scores barely reached 0.7 due to stronger clustering patterns based on cell types (e.g., alpha, beta cells) than on control/T2D labels. This conflict between cell-type clustering and label clustering explains the divergence between Silhouette (cell-type oriented) and ARI (label-aligned) metrics. Additionally, including cells with lower expressivity improved clustering by increasing

homogeneity within the latent space. Models trained on only highly expressive cells yielded similar Silhouette Scores but consistently lower ARI values.

Visualization of the latent space, seen in Figure 5, revealed that clusters were predominantly organized by cell types, with beta and alpha cells forming two distinct subclusters each. This observation motivated us to investigate whether applying the VAE to individual cell types could yield more informative latent spaces aligned with control/T2D labels.

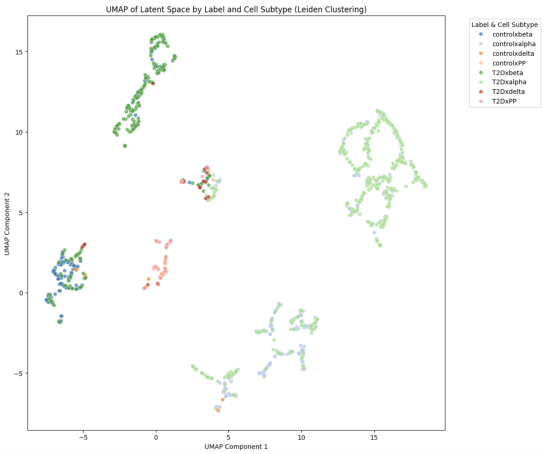

Figure 5: VAE: All Cell Types

VAE: Single Cell

Due to insufficient sample sizes, delta and PP cells were excluded from further analysis, leaving alpha and beta cells for focused VAE training. We utilized the same tools from the best model found when studying all cells and from the outset the results for single-cell-type VAEs outperformed the prior models, indicating the benefit of isolating cell types.

For alpha cells, the best models achieved a peak Silhouette Score of 0.38, but ARI and F-scores remained low, suggesting that while clusters were well-defined, they did not align strongly with control/T2D labels. In contrast, for beta cells, the VAE reached a Silhouette Score of 0.45 with an ARI of 0.25 and F-Score of 0.84, reflecting both strong clustering and improved alignment with the labels. This result is particularly meaningful from a biological standpoint, as beta cells are responsible for insulin production and secretion in pancreatic islets. Since Type 2 Diabetes (T2D) is characterized by insulin resistance and beta cell dysfunction, it makes sense that beta cells exhibit clearer distinctions between control and diabetic samples in the latent space.

The single-cell VAEs also benefited from gene selection, with the top 35% most expressive genes yielding the best results. This indicates that highly expressive genes contribute significantly to capturing biologically meaningful patterns in control/T2D distinctions within the latent space. A notable observation for beta cells was the tradeoff between latent space dimensionality and performance metrics. Higher latent dimensionality improved ARI scores but reduced Silhouette Scores, with the highest Silhouette (0.49) corresponding to a very low ARI (<0.1). This phenomenon can be explained by the fact that increasing the latent space dimensionality allows the VAE to model more nuanced and complex relationships between features, which can align better with the control/T2D labels (hence increasing ARI). However, as dimensionality increases, the clusters tend to become more spread out and less compact, reducing the Silhouette Score, which favors tight, well-separated clusters. The UMAP's of the best performing Alpha and Beta VAE's can be seen in Figure 6

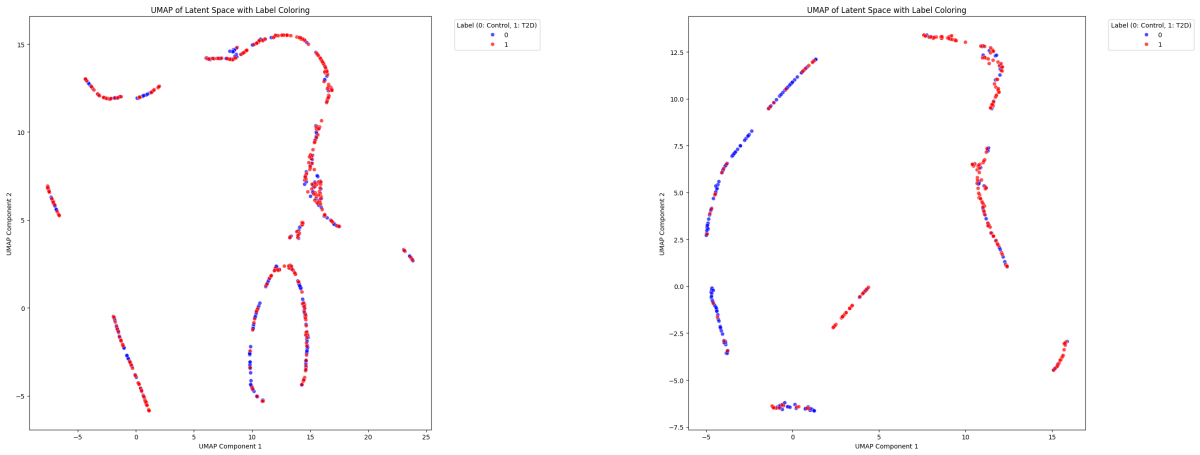

Figure 6: (Left) Best VAE for Alpha Cells; (Right) Best VAE for Beta Cells

Validation

To ensure the reliability of our results, we employed two validation strategies:

- **Random Seed Testing:** Models were trained with different random seeds to assess variability.
- **Data Splitting:** The data was split into 70/30 and 60/40 groups to test model generalization across subsets.

For all cells, the results were consistent across splits and seeds, with metrics varying within an 8% margin. For single-cell VAEs, results were slightly less consistent, varying within a 13% margin, likely due to the smaller sample size, which limits generalization. Despite these challenges, the results validate the robustness of our approach for both cell-type-specific and all-cell analyses.

Discussion

Our findings support the hypothesis that dimensionality reduction techniques, particularly VAEs, can uncover meaningful patterns in pancreatic islet cell gene expression. While PCA and t-SNE struggled to capture nonlinear relationships, the VAE latent space provided clear separation between diabetic and non-diabetic samples, especially when focusing on highly expressive genes. Additionally, UMAP demonstrated improved separation of clusters compared to t-SNE, further reinforcing the need for nonlinear methods in analyzing complex gene expression data.

The results from our VAE analyses highlight the importance of both latent space dimensionality and the biological relevance of gene expressivity:

- **All Cell Types:** While initial baseline models performed poorly, iterative parameter optimization and the introduction of techniques such as KL Warmup and ZINB gene likelihood significantly improved the quality of the latent space. Models performed better with low-expressivity genes included, likely due to their homogeneity reducing noise from strong cell-type-specific signals. However, clustering remained dominated by cell types rather than diabetic/control labels, leading to lower ARI scores despite improved Silhouette Scores. These results underscore the complexity of the dataset, where biological signals may overlap across different conditions.
- **Single-Cell Analysis:** Focusing on specific cell types, particularly alpha and beta cells, yielded more promising results. For alpha cells, VAEs achieved acceptable Silhouette Scores, although ARI and F-scores remained low. In contrast, beta cells exhibited strong performance for all scored, suggesting that beta cells are more informative for distinguishing control and T2D samples. This aligns with the fundamental role of beta cells in insulin production and secretion, which are central to Type 2 Diabetes pathology. A key observation was the impact of latent space dimensionality: increasing dimensionality improved ARI scores by allowing the model to capture nuanced, label-specific patterns, but this came at the cost of reduced Silhouette Scores due to less compact clusters. This trade-off highlights the tension between optimizing for label separation (global

patterns) and cluster tightness (local structure). In contrast with all-cells models, single-cell-type VAE's performed better when only highly expressive genes were included.

Challenges

Our study faced several challenges that impacted the generalizability and performance of the models:

- **Dataset Limitations:** The small size of the dataset (18 donors and 1600 samples) and the imbalance between control and diabetic samples introduced variability and limited the robustness of our results. This was particularly evident when analyzing delta and PP cells, where insufficient sample sizes prevented meaningful conclusions. Similarly, the Gene Id's used on the dataset seemed to be simple indexing, not translating directly to recognized Gene Id's.
- **Limited Biology Expertise:** As two computer science/statistics undergraduate students, this was our first experience interacting with Biology, let alone Genomics, topics. While this was an incredible learning opportunity that we grateful for, the project would surely have benefited from having someone with domain expertise on board.

Future Work

To enhance the robustness and applicability of our results, future efforts will focus on:

- **Validating Findings on Additional Datasets:** Expanding the analysis to larger, more balanced datasets will help validate the observed patterns and improve generalizability.
- **Refining VAE Architectures:** Further experimentation with VAE configurations, such as alternative latent space dimensionalities, regularization strategies, and incorporating biologically informed priors, could optimize latent space interpretability and improve ARI scores.
- **Cell-Type-Specific Clustering:** Integrating cell-type-specific clustering with differential expression analysis may help uncover biomarkers that are both cell-type and condition-specific, addressing the conflict between global and local clustering signals we identified.
- **Gene Subset Optimization:** Developing systematic approaches to identify the most informative gene subsets, rather than relying solely on expressivity thresholds, could further enhance the performance of both single-cell and all-cell models.

Our work can be found on the following Google Drive.

References

[1] Du, Y. and Sun, F. (2022). Hicbin: binning metagenomic contigs and recovering metagenome-assembled genomes using hi-c contact maps. *Genome Biology*, 23.

[2] Kabir, M. F., Chen, T., and Ludwig, S. A. (2023). A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction. *Healthcare Analytics*, 3:100125.

[3] Kingma, D. P. and Welling, M. (2022). Auto-encoding variational bayes.

[4] Nazaret, A., Fan, J. L., Lavallée, V.-P., Cornish, A. E., Kiseliovas, V., Masilionis, I., Chun, J., Bowman, R. L., Eisman, S. E., Wang, J., Shi, L., Levine, R. L., Mazutis, L., Blei, D., Pe'er, D., and Azizi, E. (2023). Deep generative model deciphers derailed trajectories in acute myeloid leukemia. *bioRxiv*.

[5] Xiang, R., Wang, W., Yang, L., Wang, S., Xu, C., and Chen, X. (2021). A comparison for dimensionality reduction methods of single-cell rna-seq data. *Frontiers in Genetics*, 12.

[6] Xin, Y., Kim, J., Okamoto, H., Ni, M., Wei, Y., Adler, C., Murphy, A. J., Yancopoulos, G., Lin, C., and Gromada, J. (2016). Rna sequencing of single human islet cells reveals type 2 diabetes genes. *Cell Metabolism*, Volume 24, Issue 4.