We aim to compare changes in gene expression in human islets between patients with human type 2 diabetes (T2D) and controls. The RNA sequencing dataset used in this assignment has previously been published (De Jesus *et al.*, 2019, Nature Metabolism) and is deposited with NCBI GEO under the accession code GSE120024.

Dataset. 26,463 mRNA transcripts were sequenced for 7 healthy controls and 8 patients with T2D. The age range of healthy controls varied between 26-68 years while the age range of patients varied between 47-65 years. Both groups had ~25% female subjects and 3 batches of cell culture experiments.

Data processing. m^6A sequencing data was excluded from the sample information; only RNA sequencing data was included for analysis. In this analysis, the effects of batches, sex, and age were controlled. Batches and sex were treated as categorical variables. Age was treated as a numeric variable and mean normalized to allow general linear models to converge. Sample information and the raw count matrix were checked to ensure they matched subject information. Count matrix was pre-filtered to only keep genes that have at least 10 total reads across samples. This was done to reduce the memory size of data objects and increase the speed of computation. A dds data object was created and subsequently, the data was normalized using variance stabilization (VST) for downstream differential expression analysis. VST puts the data on a log2 scale while also dealing with sampling variability of low counts. VST was chosen over rlog since it has a faster run time than rlog. Batch effect was removed by using the limma package.

Principal component analysis. PCA was performed for dimensionality reduction and data visualization. PCA creates new features called principal components, which are a linear combination of the original features. PCA suggested that batch, sex, and age do not seem to play a role in separating control and T2D (Figures 1A, 1B, and 1C) islets. PCA only slightly segregated control and T2D islets, suggesting that there might be a lot of variability in the data or that the 2 conditions cannot be separated that well using RNA sequencing (Figure 2A).

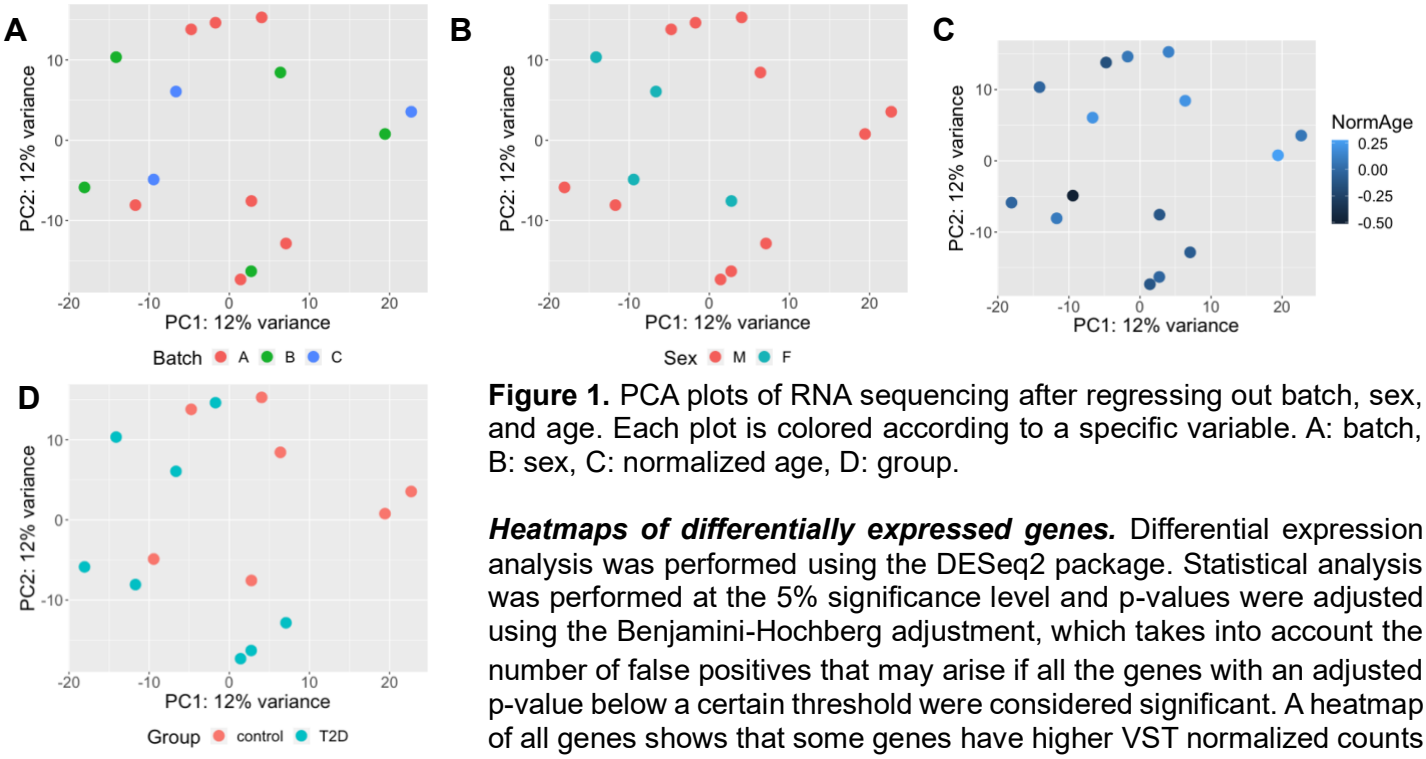

Figure 1. PCA plots of RNA sequencing after regressing out batch, sex, and age. Each plot is colored according to a specific variable. A: batch, B: sex, C: normalized age, D: group.

Heatmaps of differentially expressed genes. Differential expression analysis was performed using the DESeq2 package. Statistical analysis was performed at the 5% significance level and p-values were adjusted using the Benjamini-Hochberg adjustment, which takes into account the number of false positives that may arise if all the genes with an adjusted p-value below a certain threshold were considered significant. A heatmap of all genes shows that some genes have higher VST normalized counts than others and that control and T2D islets cannot be separated easily (Figure 2A). There were 956 statistically significant (p-adjusted < 0.05) differentially expressed genes (log2FC > 1.002 or log2FC < 0.998) between T2D and control islets. The top 10 statistically significant genes (based on p-adjusted values) showed that control and T2D islets can indeed be segregated (Figure 2B). Control islet cells have upregulation in CRABP1, CCDC169, MIR6759, and GPR149 compared to T2D islet cells. T2D islet cells

have upregulation in EPHA7, LINC00310, LINC00704, AREG, LOC283710, and LOC388780 compared to control islet cells.

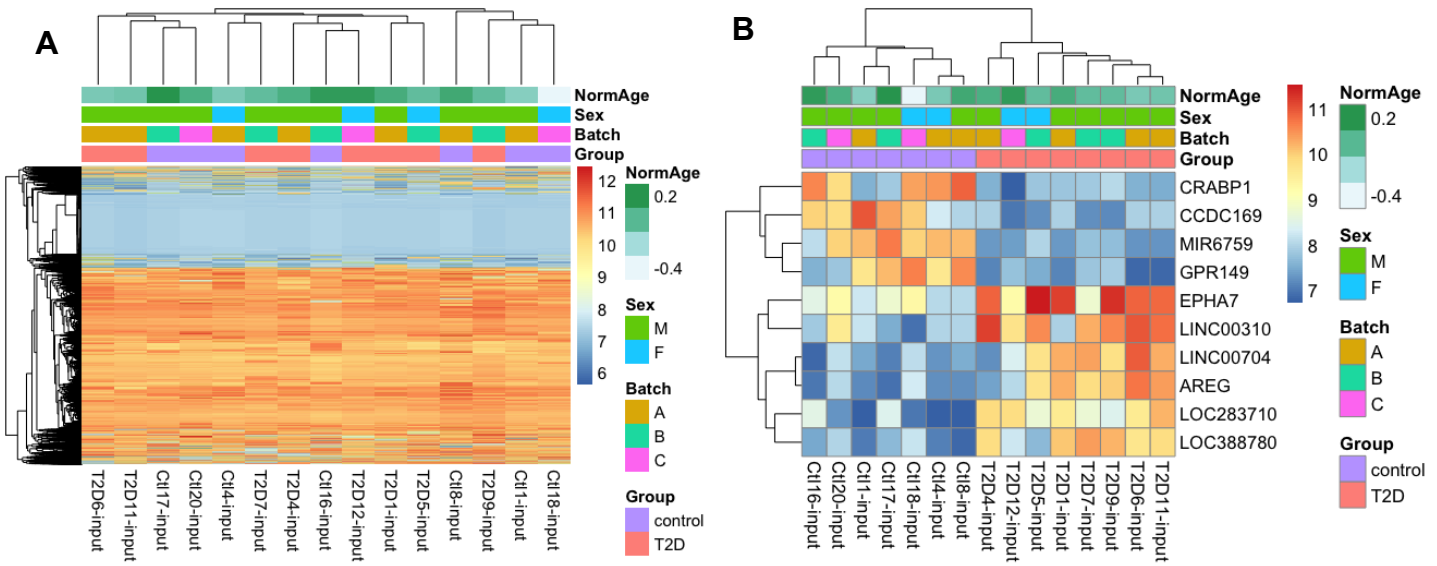

Figure 2. A. Heatmap of all genes in T2D islets compared to controls. B. Heatmap of top 10 statistically significant (based on p-adjusted values) genes in T2D islets compared to controls.

KEGG and GO pathway enrichment analysis. These approaches were used to provide functional annotations of genes and proteins. KEGG pathway analysis is used to elucidate molecular interactions and reactions. In this case, Neuroactive ligand-receptor interaction is significantly affected in T2D islet cells compared to controls. The JAK-STAT signaling pathway is also impacted; this signaling pathway regulates blood glucose levels. GO enrichment analysis provides functional annotations for genes based on biological processes. Although none of the functional annotations are statistically significant after p-value adjustment, it appears that nuclear, metabolic, and cell development processes are impacted in T2D islets compared to controls.

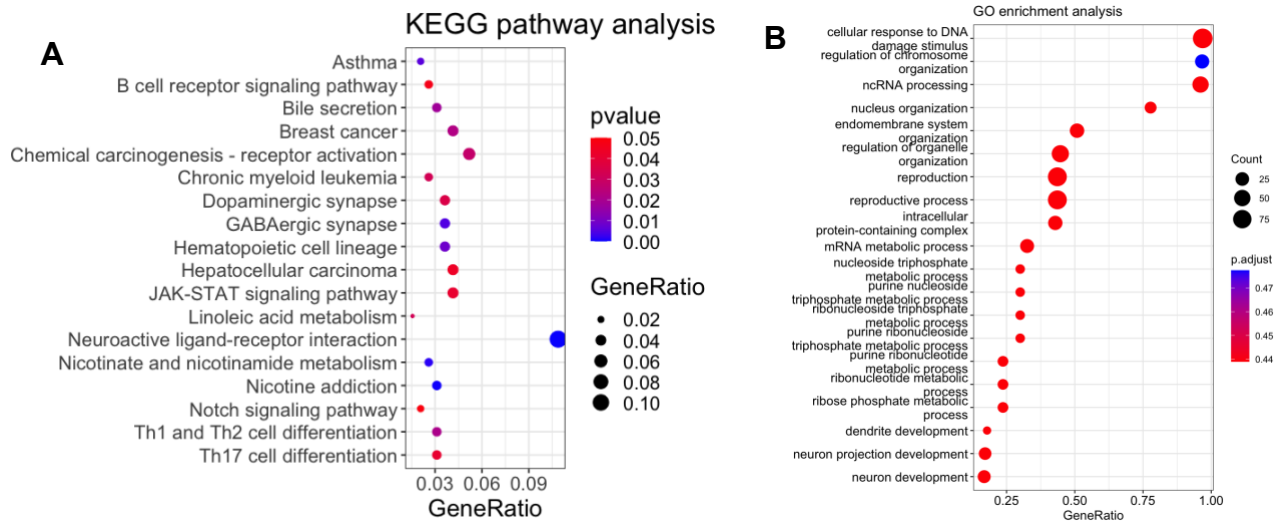

Figure 3. A. KEGG pathway annotations in T2D islets compared to controls. B. GO enrichment analysis of RNA sequenced genes in T2D islets compared to controls.

Reproducibility. R analysis codes are available at github.com/syed-adil-wafa/DGP-486-Assignment-1.