CGS4144 Bioinformatics - Assignment 2

Name	Email	GitHub
Sean Hamilton	sean.hamilton@ufl.edu	`https://github.com/sean293`
Amilcar Suarez	amilcar.suarez@ufl.edu	`https://github.com/realasuarez`
Tatum Bowen	bowent@ufl.edu	`https://github.com/tatumashley`
Joshua Roberts	robertsjoshua@ufl.edu	`https://github.com/robjob0704`
Daniel George	danielgeorge@ufl.edu	`https://github.com/danielgeorge922`

Data: www.refine.bio/expriments/SRP075806

Scientific Question: Do Obese and/or T2D myocytes show different early gene-expression response to insulin than Healthy myocytes?

GitHub Repository: `https://github.com/danielgeorge922/cgs4144_project`

---

1. When we load the expression matrix into Python, we find that the raw expression matrix has the shape $(43363, 117)$. The first number corresponds to the number of genes and the second number corresponds to the number of subjects minus the gene column. Thus, we have 116 subjects with 43363 observed genes.

   However, our matrix uses Ensembl IDs instead of gene names. To convert the Ensembl IDs to gene names, we use the Python library mygene. The resulting matrix has shape $(32059, 116)$. This means that our matrix has **116 subjects** and **32059 recorded genes**.

   To get a view of the variation within the data, we first log scale it, then create a density plot of the per-gene median expression ranges.

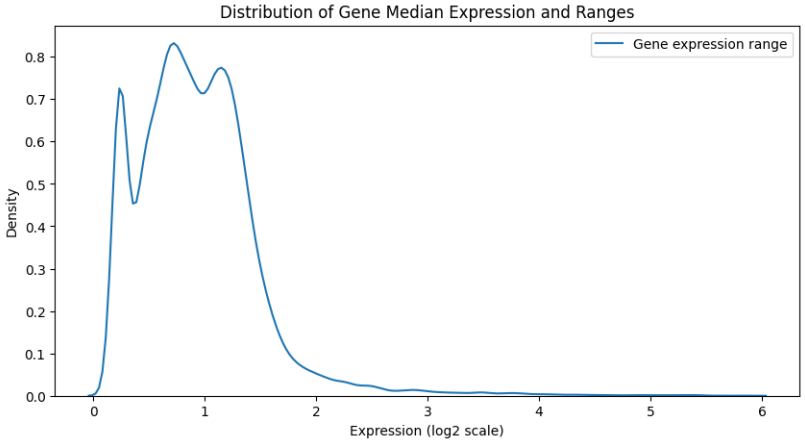

Figure 1: Distribution of gene median expression and ranges

2. Now we generate a PCA plot, a T-SNE plot, and a UMAP plot of our data.

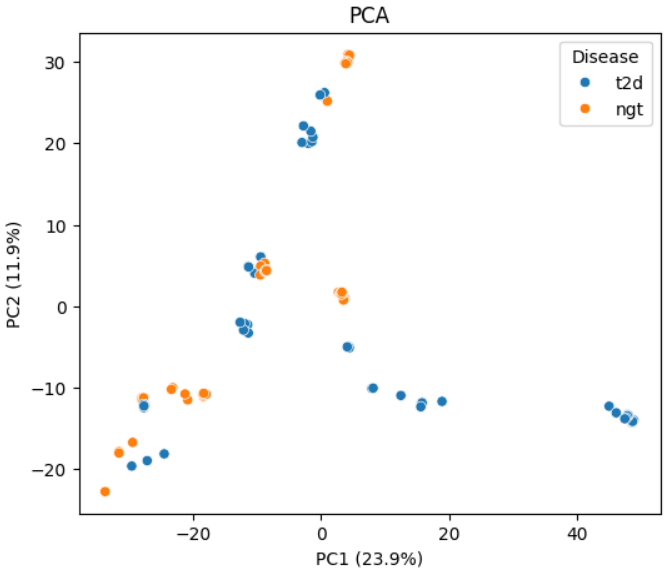

Figure 2: PCA

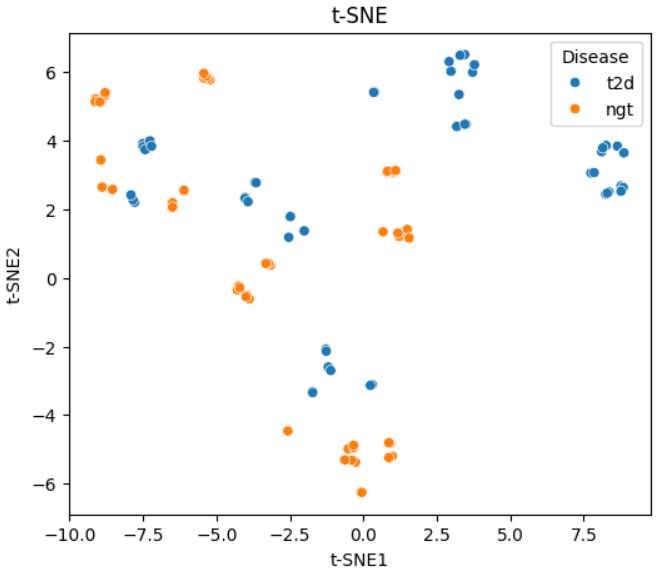

Figure 3: PCA

3. Volcano plot

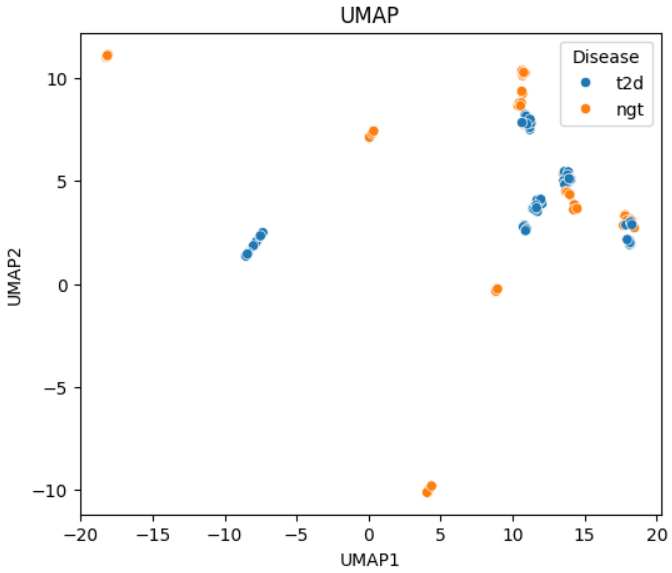

Figure 4: PCA

4. Significant genes
5. Enrichment analysis

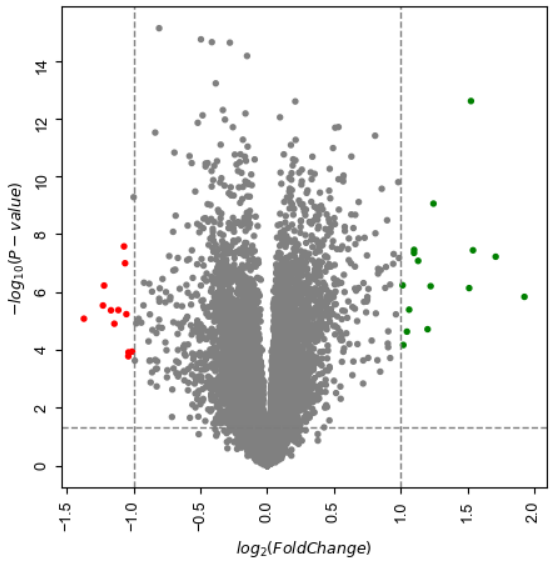
$-log_{10}(P-value)$
$log_2(FoldChange)$
0
2
4
6
8
10
12
14
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0

Figure 5: PCA