Final Project

For this project, you are required to use **Seurat version 5.0.3 in R**. Please submit 1) your ***code***, 2) ***Seurat objects*** that store the annotated cell type labels [see below for details], and 3) the ***Excel file*** with your answers to each question.

Note that you should make sure your submitted code can run without errors and reproduce the same results you submitted. Please use *set.seed(2024)* at the beginning of your code in R.

Part 1

In this part, you are given two UMI count matrices with genes as rows and cells as columns [first one: *BoneMarrow_dataset1.rds*; second one: *BoneMarrow_dataset2.rds*]. Both these datasets come from the human bone marrow tissue from a healthy donor, which is sequenced by 10X 5' technology. Each dataset may contain some or all cell types in the following list (21 cell types in total):

- Classical monocyte cell
- Non-classical monocyte cell
- Plasmacytoid dendritic cell
- Dendritic cell
- Precursor B cell
- small pre-B-II cell
- Memory B cell
- Naive B cell
- CD16 natural killer cell
- Natural killer cell
- Plasma cell
- Macrophage cell
- Megakaryocyte cell
- Myeloid progenitor
- Erythroid progenitor
- Hematopoietic stem cell
- Lymphocyte cell
- CD8 T cell
- CD4 T cell
- Regulatory T cell
- Mucosal-associated invariant T cell

However, it is unknown how many and which cell types each of the two datasets contain. You need to determine which cell types exist at your discretion. Please use the Seurat pipeline and search for cell

type marker genes to manually annotate the cell types for each dataset separately. You may search online for databases or relevant literature for cell type marker genes if needed.

Additionally, please answer the following questions and fill out the ***Excel file*** with your answers.

1. 1) What functions in Seurat did you use to annotate cell types?
   Use *Y* for Yes, and *N* for No to indicate this in the *Excel sheets Part1-Dataset1-Q1 and Part1-Dataset2-Q1*.
   2) Please record the parameter values you tried for each function in the *Excel sheets Part1-Dataset1-Q1 and Part1-Dataset2-Q1* as well.
   Note that the Excel sheet lists the parameters you are allowed to change for each function. You could try any legit values for most parameters, but for a few parameters, you can only try certain values. If you only used the default values without tuning or you didn't use the function, fill out *NA*.
2. What are the parameter values for each function you decided to use for your final cell type annotations? Fill out them in the sheets *Part1-Dataset1-Q2 and Part1-Dataset2-Q2*. [See hint 1: you should choose one way for normalization. (If you choose sctransform, then you should not use ScaleData() and FindVariableFeatures().)]
3. What cell type marker genes did you use in your attempt to identify each cell type? List the marker genes in the sheets *Part1-Q3*. Please use Ensembl IDs as marker gene names [for example, ENSG00000203747].
4. What are your final annotation results? Please add your annotated cell type label for each cell to the submitted Seurat objects as a part of the seurat.obj@meta.data, named 'cell_type'. And please use *exactly the same cell type name listed above* as your cell type labels.
   Hint: seurat_obj['cell_type'] = your_cell_type_labels

Hints:

1. Note that if you try to use sctransform as a normalization step in Seurat. It contains scaling and highly variable gene selection automatically. Thus, you don't need to run ScaleData() and FindVariableFeatures() again. But you still need to run ScaleData() and FindVariableFeatures() if you use NormalizeData(). The number of HVGs in sctransform is determined by parameter: *variable.features.n* in the SCTransform() function.
2. Ensembl ID vs gene symbol:
   Links for the explanation of the Ensembl ID
   We provide you with the Ensembl ID because of its unambiguity so that you can uniquely find the gene. Alternative gene symbols might be used in the literature to describe the same gene with upper/lower case differences.
   To find the mapping between Ensembl ID and gene symbol *for a few genes*, you could directly search online or search on the Ensembl website. Make sure you use the human database.
   To find the mapping between Ensembl ID and gene symbol *in bulk*, you could use the online biomart tools, or the biomaRt R package. Make sure you use the human database.

Part 2

In the second part, you are given one UMI count matrix with genes as rows and cells as columns [*Pancreas.rds*]. The dataset comes from the healthy donor's pancreas tissue. The human pancreas dataset usually contains ***pancreatic polypeptide (PP)*** and ***alpha cells***, which we are interested in. Please use the Seurat pipeline and search for cell type marker genes to manually annotate these two cell types *if they exist*. You may search online for databases or relevant literature for cell type marker genes if needed. Additionally, please answer the following questions and fill out the ***Excel file*** with your answers.

1. 1) What functions in Seurat did you use to annotate cell types?
   Use *Y* for Yes, and *N* for No to indicate this in the *Excel sheet Part2-Q1*.
   2) Please record the parameter values you have tried for each function in the sheet *Excel sheet Part2-Q1* as well.
   Note that the Excel sheet lists the parameters you are allowed to change for each function. You could try any legit values for most parameters, but for a few parameters, you could only try certain values. If you only used the default values without tuning or you did not use the function, fill out *NA*.
2. What are the parameter values for each function you decided to use for your final cell type annotations? Fill out them in the sheets *Part2-Q2*. [See hint 1: you should choose one way for normalization. (If you choose sctransform, then you should not use ScaleData() and FindVariableFeatures().)]
3. What cell type marker genes did you use in your attempt to identify PP and alpha cells?
   List the marker genes in the *sheet Part2-Q3*. Please use the *Ensembl IDs* as the gene names.
4. What are your final annotation results?
   Please make sure that you add your annotated cell type label to each cell in the submitted Seurat objects as a part of the seurat.obj@meta.data, named 'cell_type'. For PP cells and alpha cells you find, please use *'PP' and 'alpha' as their cell type labels*. For cells that you think are neither PP nor alpha cells, use *NA* as their cell type labels.
5. If PP cell or/and alpha cell exists, can you find new marker genes besides what you used to annotate the cell types? Fill your answers in the *sheet Part2-Q5*. If you did not find PP/alpha cells, or you don't think there are any additional cell type marker genes, please fill out *NA* in the sheet.