Introduction to single-cell transcriptomics Mandatory TD

Pick one dataset and perform the requested analyses (and more if you can). A report containing all your code as well as results and a discussion of those must be returned to me via Moodle by January 6, 2023 the latest.

Dataset 1

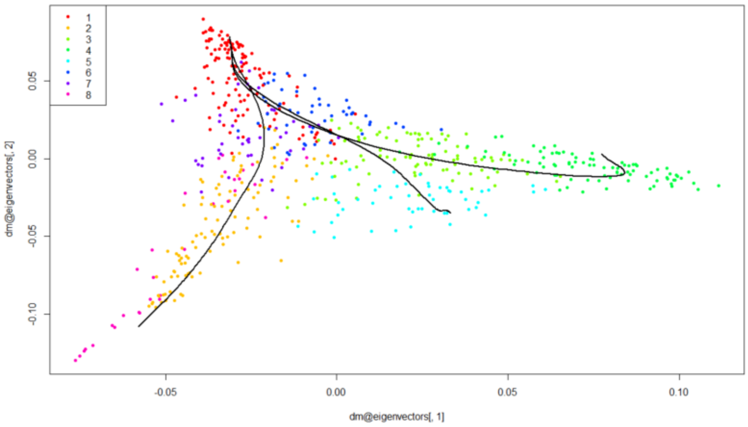
Macrophages are immune cells that have important additional functions in different tissues. For instance, microglial cells, which are the macrophages of the central nervous system, are involved in cognition beyond immunosurveillance. In tumors, so-called tumor-associated macrophages (TAMs) also have multiple roles.

In this dataset, the authors sequenced TAMs obtained from a murine model of breast cancer to explore the potential diversity of their phenotypes. That is, given their single-cell transcriptomes, we want to identify clusters, check whether they express different transcriptional programs although they are all TAMs, and to postulate potential roles based on the specific genes of each cluster.

Read the data with Seurat (files `barcodes.tsv.gz`, `features.tsv.gz`, and `matrix.mtx.gz`), keep genes expressed by at least 10 cells and cells with at least 800 distinct genes fulfilling this condition. Then, eliminate cells with >12% mitochondrial genes and cells with >65,000 total UMIs.

As usual, identify the 2,000 most variable genes and, then, eliminate from this list all the genes involved in the cell cycle (GO term Cell Cycle for the mouse). You should be left with roughly 1,660 genes. Perform PCA and identify 8 clusters by setting the resolution parameter of `findCluster()` at 1.5 and using the 35 first principal components. Next, find the marker genes and set a threshold on the adjusted P-values, *e.g.*, 10^{-2}, to obtain a list of differential genes specific to each cluster. Use any GO enrichment tool to characterize the 8 clusters.

As you can observe in the UMAP or t-SNE projection, the clusters are not well-separated. This makes sense since we are dealing with a single population of cells (macrophages). Differences are cell states, which is much less than different populations (this is the reason why we had to increase the resolution parameter for the cluster-finding function). Now, we want to infer how macrophages present in a tumor might get differentially activated to achieve different states and phenotypes. For this, we want to compute diffusion maps (package "destiny") from the 35D projection of the data after PCA (obtained from the Seurat object with `Embeddings()`). The diffusion map is an object that contains a projection as well, you can plot using the first two dimensions (slot eigenvectors in the dm object). Lastly, you can use the 15 first dimensions of the DM to apply the library "slingshot" and predict trajectories. Use as a starting point the cluster that had the least GO term enrichment, *i.e.*, the least differentiated TAM subpopulation. You should get something like this:



Dataset 2

Cancer-associated fibroblasts (CAFs) play an important role in the development of many tumors by secreting different soluble factors stimulating growth, vascularization, and inducing an immunosuppressive environment. They also induce a desmoplastic reaction remodeling the extracellular matrix.

This second dataset contains CAF transcriptomes from colorectal liver metastases (CRC-LM). Start by loading the expression matrix with `load("caf-data.rda")` to get a matrix called `caf.data`. This matrix can be given to `CreateSeuratObject()`. Keep cells with at least 1000 distinct genes detected in 1% of the cells. Then eliminate cells with >50,000 total UMIs and >50% mitochondrial genes.

Perform the classical clustering and projections to obtain 4 clusters. Use GO term enrichment to characterize the phenotypes of the different CAF subpopulations. To fine the differentially expressed genes across clusters, use the `FindAllMarkers()` function.

CAF may originate from different sources. Obviously, resident fibroblasts but also mesenchymal cells. In the liver, resident fibroblasts activated by cirrhosis are called SAMes. They are a realistic model of activated, resident fibroblasts. Vascular smooth muscle cells (VSMC) and hepatic stellate cells (HSC) are two other mesenchymal populations likely to differentiate as CAFs. Based on the gene signatures in the table below, try to predict the most likely origin of each CRC-LM CAF subpopulation.

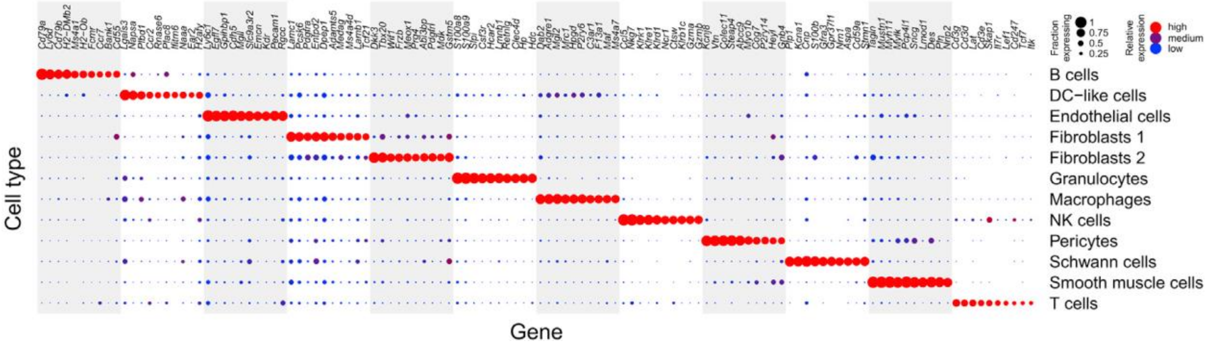
	VSMC	HSC	SAMes
1	PLN	IGLC2	LUM
2	SORBS2	HIGD1B	PTGDS
3	PHLDA2	IGHA2	IGFBP3
4	SNCG	RGS5	COL1A1
5	MT1M	TM4SF1	COL3A1
6	RP5-966M1.6	FABP5	DPT
7	MYH11	PLAT	COL1A2
8	ACTG2	FABP4	DCN
9	ADIRF	AGT	FBLN1
10	CRIP1	CPE	CCL19
11	ITGA8	NPR3	C3
12	PDGFA	IGHA1	C7
13	SBSPON	SSTR2	SERPINF1
14	PPP1CB	FCN3	TIMP1
15	AC097724.3	AC018647.3	CXCL12
16	KCNMB1	IMPA2	LXN

Dataset 3

Mouse heart muscle cells from two male and two female animals were sequenced. Read the expression matrix from `full_count_matrix.txt`. The gene symbols are in column 1 meaning that you have to define the rownames accordingly and then suppress this column before creating the Seurat object.

Keep cells with at least 500 distinct genes detected in 0.5% of the cells. Then eliminate cells with >18,000 total UMIs and >12% mitochondrial genes. Perform dimension reduction and clustering as usual, and aim at obtaining roughly 12 clusters.

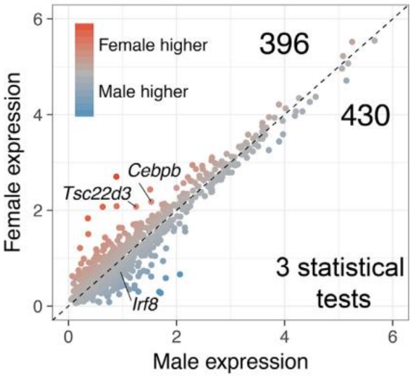
Call the cell types of (most of) your clusters. The figure below might help for this:



For each cell population, search for sexual dimorphism. To do this, you have to classify each cell as male or female. You can use the following rule:

- If *Xist* is expressed and none of the Y-chromosome genes *Ddx3y, if2s3y, Erdr1, Gm29650, Kdm5d,* and *Uty*, then the cell is classified as female;
- If *Xist* is not expressed, but the Y-chromosome genes *Ddx3y, if2s3y, Erdr1, Gm29650, Kdm5d*, and *Uty* are expressed, then the cell is classified as male;
- Unclassified otherwise.

Unclassified cells are ignored. You can use `FindAllMarkers()` to search for differential genes between males and females, provided you defined new Seurat objects with just one cell population at a time and as "clusters" the "Idents" male of female (unclassified cells should not be included in such new objects). A plot for each cell type comparing the average expression level of genes in male and female cells with differential genes labelled must be generated; example below.



Algorithm 1

Use the PBMC data and SingleCellSignalR to infer ligand-receptor interactions (LRIs) between cell populations (*cf.* `SingleCellSignalR-template.R`). Write a R function to generate a new graphical representation of all the LRIs after the following template:

Outgoing **Incoming**

Ligand	CP$_1$	CP$_2$	…	CP$_n$
L$_1$	░	▓		
L$_2$				▒
L$_3$		▒		
L$_4$	▓	█		
…				

Receptor	CP$_1$	CP$_2$	…	CP$_n$
R$_1$				▓
R$_2$	▓			
R$_3$		█		
R$_4$	░	░		█
…				

In the above example, cell populations 1 (CP$_1$) and 2 secrete a ligand L$_1$ that interacts with a receptor R$_1$ secreted by CP$_n$, etc. The color-scale must reflect how much of the total LRI scores (LR-score in SingleCellSignalR package) is represented by a given interaction L$_i$-R$_j$ between CP$_k$ and CP$_l$. This value can be computed by dividing the LR-score of this specific interaction by the total of all the predicted interaction LR-scores between all the cell populations.

The output must be a plot and you might want to use ComplexHeatmap for this purpose. Putting all the LRIs from one dataset in a single plot will result in an image with a huge height. Hence, add the functionality to limit the plot to user-supplied lists of ligands, receptors, and cell populations. In addition, you could offer the possibility to specify a certain GO term or Reactome pathway to which the plot should be restricted. Data frames associating gene symbols to GO terms and Reactome pathways are available from SingleCellSignalR directly, but you could also consider other sources.