Single cell analysis with R/Bioconductor hands on session

In this session we are going to use the Orchestrating Single Cell Analysis (OSCA) resource to get some experience of data analysis in the context of scRNA-seq data:

https://bioconductor.org/books/release/OSCA/book-contents.html

An alternative resource is this tutorial:
https://www.singlecellcourse.org/index.html
We will not follow this but would refer to this occasionally.

In RStudio or your favourite IDE, open the R script I have emailed you and run it to install the required packages and also load the relevant datasets. You can see more details of how the data is loaded in the following OSCA workflows. In the following, copy and add the relevant code from the different parts of the OSCA book to this script and run these line by line.

Start with the OSCA Introduction chapter 4, which introduces the SingleCellExperiment (SCE) class which is a common data infrastructure for scRNA-seq storage. Figure 4.1 gives a good overview of the data structure. You could either just skim through this chapter (without running the code) and familiarise yourself with SCE data structure.

[If you want to run some of the code, you need to download the dataset that is used in this section from the following link:
https://www.ebi.ac.uk/biostudies/arrayexpress/studies?query=E-MTAB-5522
Using one of the counts_Calero files for the main count data (section 4.3.1) and for meta data using the .sdrf file (section 4.4.1).]

[More details on R data types and classes and introduction to SCE class can also be found in the alternative tutorial mentioned above chapters 4 and 5. Here, you can learn that SCE is a class from S4 system, which is R's formal object-oriented programming system.]

Next move to the Basics book:
https://bioconductor.org/books/3.22/OSCA.basic/index.html

1. Quality Control Skip section 1 on Quality Control or skim through it quickly (we discussed this a bit in the lecture).

2. Normalisation. The required dataset from workflow 2 is already loaded using the script provided. Go through 2.1, 2.2, 2.3 and 2.5 (feel free to skip 2.4). Note that in log transforming a pseudo-count of 1 is added (you can look at OSCA Advanced book section 2 for more details on normalisation and transformations).

3. Feature Selection. The 416b dataset and PBMC dataset are loaded using the script following Workflows in Chapter 1 and Chapter 3 . Try 3.2 and 3.3. Do a scatter plot of the biological noise based on spike and based on the trend of the data (part 3.2). Are they always equal. Compare the top HVGs. (3.4 optional, don't do 3.5 and 3.6).

Let's continue with more advanced material from the OSCA Advanced section 3.4.

4. Dimensionality Reduction. The dataset is already available. Do all the sections, try PCA on different number of HVGs, do you notice a difference in the results. Try UMAP and t-SNE on more coordinates, which one is more efficient?

More advance topics on choosing the number of PCA components is reviewed in the OSCA Advanced section 4.2. Also, some more efficient methods of visualisation are discussed in 4.4 that you may consider to review.

5. Clustering. The dataset is already available. Go through 5.1 to 5.4, you can skip 5.5. the OSCA Advanced section 5.1 to 5.4 provide some more in depth analysis for testing cluster behaviour and stability (using bootstrapping). There is a lot here so you are welcome to explore as much as possible.

6. Marker gene detection and Cell type annotation. For those who are interested you can go through 6.1 and 6.2 and 7.1 and 7.2 for a taster on these more advanced topics for biological interpretation of the data.

7. Trajectory Analysis. This topic from the OSCA Advanced is indeed an advanced topic for the interested. The dataset is sce.nest from workflow 10 that can be downloaded using the R script provided. The chapter goes through a simole method baed on minimum spanning trees and also RNA-velocity for trajectory inference.