Extra assignment

Instructions

Answer these questions and send them to one of the TAs assigned to the course, the deadline is the same as for the standard computer lab. For the subject line in your email use "[CourseCode]-FirstNameLastName", where the CourseCode is either CD2040 or BB2255. To avoid getting your hand in marked as "late" by canvas, hand in an empty html file named "blank.html" and write in the comments that you've sent the report to the email given above. Your answers should be clear and longer than just a sentence, also make sure that you motivate your reasoning thoroughly.

Questions

Q1.

Working with both single cell and spatial transcriptomics data, it's a common procedure to apply some form of transformation to the data in order to normalize it. For example, in Lab2 and 3 this is done by using the SCTransform function provided by Seurat (an R package). Why is such a procedure used, i.e. what artifacts are we trying to eliminate upon normalizing our data?

Q2.

Working with ST data, we sometimes need to remove spots and genes of low quality - a process referred to as filtering. Explain :

(a) Why we tend to remove genes only occurring in very few spots?
(b) Why spots with very low total amount of genes, or very few unique genes are usually excluded from our analysis?

Q3.

Working with biological data a common experimental design is to have a case and control group respectively, or in more general terms a steady vs. perturbed system. A question that naturally arise from such a design is what the differences between the two are (eg. between case and control). Working with transcriptomics data, this question may be posed as "what genes are over-expressed in the case compared to the control". Answering these questions could potentially aid us in associating certain genes to a given condition. To perform such a study, once the data has been collected, we conduct what is known as a "Differential Expression Analysis" (DEA). Two standard metrics that are obtained from DEA are p-values and Fold changes (usually given as log2-values). Please explain :

(a) The information that the p-value contains in the context of DEA.

(b) The information that the log2 Fold Change contains in the context of DEA.
(c) Why we often obtain both a p-value and something referred to as an " adjusted pvalue". What are we adjusting for?

Q4.

Analyzing ST and Single Cell data, we are usually interested in how cells or spots group together, i.e. what natural clusters that are found within the data. One of the simplest algorithms to cluster data in an unsupervised manner is "Kmeans", which is nothing else but a special case of a GMM (Gaussian Mixture Model) with hard labels. Kmeans basically works by iteratively assigning data points to a cluster based on the distance to the centroids of said clusters - a data point is assigned to its nearest centroid - until convergence is reached.

(a) How do we measure distance (not physical) between spots? Give two examples of how we could compute the "distance" between spots in the expression space.
(b) Why do we use techniques such as PCA and ICA prior to clustering our expression data? Your answer should involve some reference to the "curse of dimensionality".

Q5.

Assume that you are given a set of ST data, taken from a melanoma patient. The spots in your data cover both malignant and benign tissue. There's a large corpus of scientific material regarding this disease and you find two genes of interest upon searching within the literature:

- Gene A – associated to the tumor proliferation.
- Gene B – suggested to be a key component of the tumor micro environment (the area adjacent to the periphery of the tumor regions).

Briefly describe how you can use the spatial data to quickly assess whether the presumed properties of these genes seem to be correct or not.