Neuro-Genomics
Final project - Sequencing Data Analysis

General instructions for this project:

a) Describe your workflow in analyzing the raw data. Use a text file for that (Word or a similar text editor) and upload it as part of your project. Which steps did you take? What exactly was performed in each step? How did you reach your conclusion about the possible molecular deficiencies in the neurological disease (Part 1), and the outcome of the cancer drug (Part 2)? Be very clear about the steps that you took all the way from the raw data to the conclusions.
b) If you are using a software in one or several steps, justify the use of the software - how do you know that this software is appropriate for the task at hand?. Write all the commands you used to run the software.
c) If you are writing your own scripts in one or several steps, make sure that it is running properly and that your code is clearly explained with comments inside the code. We need to be able to read the code and understand it. Upload the scripts with your project.
d) The code for this exercise can be written either in R, Matlab or Python.
e) If you created additional text files as part of your workflow (for example, processed data), upload them with your project.

PART 1 - analysis of bulk sequencing

In this part of the project you will help a researcher in his/her quest to understand the molecular mechanism of a neurological disease. The patients of this disease have unique physiological deficiencies which are manifested in severe neurological symptoms. Several years ago it was discovered that in patients of this disease there is mutation which prevents the translation of a specific protein. This gene is not mutated in the general population. However, the link between this gene and the physiological aspects of this disease remained unclear.

To gain molecular insights about the disease, the researcher created a **mouse** model with a knockout of the specific gene (i.e. the protein can not be created). He/she then performed a bulk sequencing experiment using cortex tissues. Three cortex tissues were sequenced from three normal control mice (marked ‘C’), and three cortex tissues were sequencing from three mice that had a knockout for the specific gene (marked ‘KO’).

Your job in this project is to analyze the raw sequencing data that the researcher created, and to describe the molecular deficiencies which are associated with the knockout gene. This can shed light into the molecular mechanism of this neurological disease, and might even suggest a treatment direction.

The raw data

As usual, the RNA extracted from the cortex tissue was fragmented during the library preparation, i.e. before the actual sequencing. The estimated average fragment length was 300 bases, and the estimated standard deviation of fragment length was 50 bases. The sequencing

reads generated were single-end, i.e. these are not paired-end reads. This information might be usual for you, depending on the software that you will choose to use.
The raw fastq sequencing files are given below.
Note: to save disk space you might want to download only one file at a time, process it, and then delete it before downloading the next file.

C1.fastq
C2.fastq
C3.fastq
KO1.fastq
KO2.fastq
KO3.fastq

If you are planning to use Kallisto or Salmon read these notes:

1. For this project it is better to work on the level of the genes and not on the level of the transcripts (here 'transcript' means the different isoforms that can be expressed from each gene due to alternative splicing). As we discussed in class, Kallisto and Salmon are based on transcript-level quantification. Therefore you need to sum all the **estimated counts** from the transcripts of an individual gene. In class we showed how to sum the TPM of all the transcripts of an individual gene, so you can use that script as a starting point.
2. After you get the estimated counts for all the genes, round these expression values to the nearest integer. This is because most differentially expressed tools expect integers to correctly model the count (expression) data.
3. Instead of (1) and (2) above, you can use the R package tximport to convert estimated counts from the level of transcripts into rounded counts in the level of the genes. Using 'tximport' is more accurate but it is not mandatory for this project -- the method described in (1) and (2) above can be used instead and it will be good enough.
4. If you prefer working on the level of transcripts (instead of on the level of genes), consider using sleuth for differential expression analysis.

If you are planning to use Bowtie2 or a similar general aligner software read this note:
For this project it is better to work on the level of the genes and not on the level of the transcripts (here 'transcript' means the different isoforms that can be expressed from each gene due to alternative splicing). Therefore consider using the RefSeq genes as a reference, instead of aligning against the genome or aligning against the all transcripts. Note that you will need to create indexes for the genes, for example using bowtie2-build.

If you are planning to use DeSeq2 for differential expression analysis read these notes:

1. Recall that the input of DeSeq2 is data.frame.
   Note that the R function 'read.table' can convert a text file into a data.frame.
   If the first line (row) of the text file contains one less entity compared to the rest of the file, the first line of the text file is interpreted as the columns names of the data.frame,

and the first column in the rest of the text file is interpreted as the rows names.
For example:

	C1	C2	C3	KO1	KO2	KO3
GeneT	17	13	6	9	3	7
GeneZ	9	5	8	12	4	2

Will be read as data.frame with columns: C1, C2, C3, KO1, KO2, KO3, and rows: GeneT, GeneZ.

2. DeSeq2 expects as an input the raw count data. Therefore, do not perform between-samples normalization, or any other normalization procedure, on the count data which is the input of DeSeq2.
3. DeSeq2 requires information about the conditions provided in the counts data.frame. Specifically, information is needed about which conditions are 'untreated' (in our case these are 'C1', 'C2', C3' = the control tissues), and which conditions are 'treated' (in our case these are 'KO1', 'KO2', and 'KO3' = the knockout tissues).
   For your convenience, we provide a conditions file which looks as follows:

condition	type	
C1	untreated	single-read
C2	untreated	single-read
C3	untreated	single-read
KO1	treated	single-read
KO2	treated	single-read
KO3	treated	single-read

   Make sure that the columns in your counts file fit both the order and the names in the first column of the conditions file above (i.e. 'C1', 'C2',...,'KO3'). If this is not the case, either change the order and/or names in the conditions file, or change the order and/or names in the counts file.
4. For your convenience, these are the steps to run DeSeq2 in R:

```
library("DESeq2")
# change PATH below to the location in your computer
cts<-read.table(file="PATH/EcountsPerGene.txt", sep = '\t') # the counts file
coldata<-read.table(file="PATH/conditions_file.txt", sep = '\t') # the conditions file
dds <- DESeqDataSetFromMatrix(countData = cts,
                              colData = coldata,
                              design = ~ condition)
dds <- DESeq(dds)
res <- results(dds)
```

5. Note that the log2 fold change of DeSeq2 is defined as condition untreated vs treated. This means that positive values of log2 fold change represent genes whose expression were higher in the untreated compared to the treated. For example, a gene with a log2 fold change of 1 means that the gene expression was two fold (x2) higher in the untreated compared to the treated. Likewise, a gene with a log2 fold change of -1 means the gene expression was two fold (x2) higher in the treated compared to the untreated.

6. To explore the results, consider using the R function ‘subset’. This function can be useful in focusing on genes with low p-value (i.e. differentially expressed genes). This can also be useful in focusing only on the genes that had higher expression in the control conditions compared to the knockout or vice versa (see below).
7. To export the results, consider using the R function ‘write.table’.

Describe the molecular deficiencies in the disease

To describe the molecular deficiencies in the knockout conditions, first detect genes that are differentially expressed between the control and the knockout. A common p-value cutoff to detect differentially expressed genes is p-adj (i.e. p-value adjusted for multiple testing correction) which is less than 0.01-0.1, for example, p-adj<0.01, or p-adj<0.05, or p-adj<0.1. Manually examine the genes detected as differentially expressed with the highest statistical significance (i.e. lowest p-value).

Study **separately**: (a) the group of genes that were lower in expression in the treated conditions (i.e. genes that had higher expression in the control conditions compared to the knockout), and (b) the group of genes that had higher expression in the knockout conditions compared to the control.

Can you detect something common among the functions of these genes? GeneCards is a good place to find information about the function of individual genes.

Next, perform functional analysis on all the genes that were detected as differentially expressed in groups (a) and (b) above (perform separate analysis for group (a) and group (b)). Describe the molecular functions most affected by the knockout.

Given the molecular functions affected, what is the main deficiency that you detect in this disease? Can you suggest a possible treatment direction?

PART 2 - analysis of single cell sequencing in situ

In this part of the project you will analyze a biopsy from a breast cancer patient. The cells in the tissue biopsy were sequenced using a single cell technology which preserves the location of the cells inside the tissue (i.e. in situ sequencing). Your goal in this part of the project is to predict if this individual patient will benefit from an immunotherapy drug.

Background

Immunotherapy, and especially checkpoint inhibitor therapy (Nobel prize for Medicine 2018), is a game changer in cancer therapy, saving the lives of many patients with several types of cancer. However, in some cancer types, such as breast cancer, immunotherapy is only beneficial to some patients and not all. The mechanism of action for immunotherapy is to make immune cells, and particularly T cells, less selective so that they can attack tumor cells. Therefore, for immunotherapy to work immune cells should be present in the tissue next to the tumor cells. The checkpoint inhibitor drugs which are approved for breast cancer were designed to target and inhibit the molecule PD-L1. Therefore if there is no expression of PD-L1 in the biopsy, immunotherapy is not likely to work.

The raw data

The first file contains the expression values (count data) for 291 genes in 8627 single cells that were sequenced from one patient biopsy. Each cell has a representative number. In the second file, the location for each one of the sequenced cells inside the biopsy is provided (using the same representative numbers). The location for each cell is one point in space - representing the middle of the cell. The third file contains the marker genes of each cell type.

expression_matrix.csv
locations_of_cells.csv
marker_genes.csv

Predict if the checkpoint inhibitor drug will work for the patient

Analyze the single cell data and determine the cell type of each cell.
The list of 291 genes includes known cell type marker genes, and this can help you identify the cell type of each cell.

Next address the three questions below:

1. Does the sample contain at least 10% immune cells (from the total number of cells studied)? The immune cells are T cells, Macrophage, and B cells.
2. Are the immune cells mixed with the tumor cells (MBC) in the biopsy? The alternative is that the immune cells reside in one location inside the biopsy and the tumor cells reside in a spatially different location. The answer to this question can be qualitative. For example, you can use plots to justify your answer.
3. Are at least 10% of the cells in the biopsy express the gene for PD-L1? Note that PD-L1 is not the official gene symbol name for this gene. For this question even a count of 1 for this gene in a cell should be considered as gene expression.

If the answer to all three questions is ‘yes’, then there is a good chance that the individual will respond to an immunotherapy drug from the PD-L1 inhibitor family.