# Introduction to cancer genomics exercisesfall 2018 – day 4

For today's exercises, we want to study the skin melanoma cases from TCGA. We are interested by the difference in immune infiltration between the samples. In particular, we want to observe if samples taken from lymph nodes are more infiltrated than samples directly from primary tumors.

In order to study this, we will use gene expression data and some clinical data indicating the tumor site of the samples.

There are multiple websites from where we could get such data, given in various formats. For this exercises, we will use the data found on Synapse website, which contains well annotated clinical data:

<https://www.synapse.org/#!Synapse:syn1910208>

In order to download these data, you need to register for an account on this website (it's free). To simplify your work, we provided for you the data at: <https://drive.switch.ch/index.php/s/uKCnB6AA6kfaTp5>.

Today, we also provide a script to analyze this data. Similarly to day 3 exercises, most of the code is already given in this script and it includes some explanations about the functions we used and why / how. Even if many lines of codes are already provided, try understanding what they are doing and ask us if you need help!

There are also some lines of codes that are missing (written with *### MISSING CODE ###*) and that you need to complete. If you don't find how to do it, try looking back at some code from day 3, which contained some of these "missing lines", or if still no idea, ask us for help.

## Preparation

Create a directory for today's exercise (e.g. exersiceDay4) and copy the provided script "*code\_intro\_cancer\_genomics\_day4.R*" into this directory. Create also a subdirectory called "*Data*" containing the various data files provided. The most important files are

 "*unc.edu\_SKCM\_IlluminaHiSeq\_RNASeqV2.geneExp.tsv*" (the gene expression data found on Synapse website) and "*nationwidechildrens.org\_SKCM\_bio.patient.tsv*" (the information about each sample, also found on Synapse website). The file "*TCGA\_SKCM\_primary\_tumors.txt*" could be used for exercise 4 and the two "….RData" files are there in case you cannot install some tools for the last exercise.

With help of the script, import this data into *R*, creating a matrix of the normalized counts for each gene in each sample and another data.frame with the clinical information corresponding to each sample from the gene expression data.

## Exercise 1

Draw a simple plot showing the correlation between the genes *CD3E* and *LCK* in these samples.

## Exercise 2

First draw a simple histogram for some gene of your interest (e.g. "MLANA") in these samples.

*Question:* Determine then which are the different possible tumor sites for the samples in this melanoma data?

*Note:* the tumor site of origin of each sample (i.e. primary tumor, lymph node, …) is indicated in the clinical data table. It is given in the column "*submitted\_tumor\_site*".

Then, draw a boxplot of the expression of your above gene of interest in function of the tumor site of the sample.

For the rest of the analysis, we will only keep the samples that are from *"Primary Tumor*" and "*Regional Lymph Node*". Modify the *samples.table* and *normCounts* variable to only keep these samples.

## Exercise 3

We will first look for the difference in immune infiltration in a simple way. In fact, we will already observe significant differences with this analysis. We will then nevertheless also use other more refined techniques that could be useful for more elaborate analyses.

Using boxplots, show the expression of various immune-related genes in the samples coming from *primary tumors* vs. those samples from *regional lymph nodes* (you can take for example the expression of *CD45* (synonym name: *PTPRC*)*, LCK, CD3E* and *CD19*). Are the observed differences significant between the two groups?

## Exercise 4

Next, using our in-house developed package EPIC, estimate the fraction of various immune cells infiltrating each *primary tumor* sample. In this exercise, we will use the web-application of EPIC that is available at: <http://epic.gfellerlab.org>

You first need to export from R the matrix of the gene expression from each primary tumor sample (rows need to correspond the genes and columns need to correspond the samples in this file).

*Note:* this exported file is also already present in the *Data* folder from today's exercises in case you want to try this web application without doing first the previous exercises.

 Import then this file into the web-app, to use it as the bulk sample data. And run EPIC, keeping the default values for all the other parameters.

*Note:* in EPIC, what is called *"Other cells"* correspond to the cell types that are not included in the gene expression reference profiles (i.e. this is mainly the cancer cells in tumor samples).

Which cell types are included in the predictions? What are the two most abundant cell types in average? (*hint:* look at the different figures obtained by the web-application).

## Exercise 5

If you need to analyze repeteadly such data with EPIC or want to analyze the results grouped in different categories, it can be more convenient to use the package we developed for R. Install it on your computed (see details in the script from the exercises). Then, estimate the various cell types fractions present in all samples and verify if some of these cell types are more abundant in the lymph node samples than in the primary tumors.

## Exercise 6

In this exercise, we will consider gene signatures of the immune system and perform single sample gene set enrichment analysis (ssGSEA) related to these signature for each sample.

We first need to obtain the gene signatures we are interested: the *IMMUNE\_SYSTEM* and *ADAPTIVE\_IMMUNE\_SYSTEM* gene sets from REACTOME (a list of many gene sets is available on the website of the Molecular Signature Database: <http://software.broadinstitute.org/gsea/msigdb>, where you can also find which genes are present in each signature; e.g. for the immune\_system from Reactome: <http://software.broadinstitute.org/gsea/msigdb/cards/REACTOME_IMMUNE_SYSTEM.html>).

With help of the GSVA package from R, compute the ssGSEA for the genes from these gene sets and do a similar plot than above: a boxplot showing the difference for these two scores in the primary tumors vs. regional lymph nodes.

*Note:* we computed here ssGSEA scores. This gives a score based on a gene signature for each sample independently of each other. If you want to test if a given gene set is significantly enriched between samples from one condition vs. samples from another conditionyou could make use of GSEA. These two methods are similar but don't ask exactly the same question. In ssGSEA, we usually compute only the score from specific gene signatures for each sample and compare these scores between the samples. On the other hand, in GSEA we'd first rank the genes based on the differential expression between the two conditions tested (i.e. we'd have a single ranked gene list instead of one ranked gene list per sample). We'd then compute the scores for many different gene signatures and use statistical tests to find out which gene signatures are best discriminating the two conditions tested.

You can find more information about the GSEA there:

<http://software.broadinstitute.org/gsea/index.jsp> (you can download a free software from there to to do such analysis with some graphical user interface – i.e. no need of coding yourselves).