# Introduction to cancer genomics exercises fall 2018 – day 3

In this exercise we will analyze RNA-Seq data of human peripheral blood cell populations. We will use the data published by Pabst C *et al*. (<https://www.ncbi.nlm.nih.gov/pubmed/26834243>) that can be downloaded from NCBI GEO (Gene Expression Omnibus) repository (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE51984>).

Have a quick glance at this GEO page (no need to read in details all the information, but try to have an overview of what information is given there). Click also on some sample names found in this page to get more details about these samples.

To what do these samples correspond? Whan antibodies did they use to sort the cells?

Instead of analyzing raw sequence reads files, we will start working with pre-processed files of read counts mapped to each gene. You can find them in the *GSE51984\_RAW.tar* ‘Supplementary File’ from the GEO page.

Download this file, and expand it. You should find 24 compressed files, one for each sample. Uncompress them and you should be able to view them in any text editor.

Create a directory for this exercise (e.g. exerciseDay3) and move into it the *GSE51984\_RAW* directory containing the 24 text files.

Compare two random files and try to interpret the meaning of the different columns. Determine which sample correspond to what cell type and donor, based on the sample names.

*Hint*: you can also find the definition of each column from these Supplementary files and the description of the sample if you click on a sample name in the GEO Accession viewer from above (e.g. by clicking on the link to sample *GSM1256809*) and then reading the "Characteristics" and "Data processing" sections.

We provide a script to analyze this dataset. It is organized in 4 parts and commented. Most of the code is already given, but there are some questions and small problems along the code (marked in the code with *###Q*?).

The audience of this course is diverse and some will therefore be faster at doing all the exercises than others. The parts 1 and 2 should be done first as they introduce basic concepts/functions used in the other parts and tomorrow’s exercises. Then, depending on the time remaining and your interest, you could decide to do the part 3 or 4 (you can come back to the other part later on as all the necessary information for these analyses is already present in the script).

With the Part 1 of the code (marked with "*# Part 1 -----*" in the code) you should be able to import the data into *R* (both raw counts and RPKM normalized expression values) and generate plots (histograms, boxplots) for the expression of specific genes (e.g. CD4, LCK) in different cell types (e.g. B-cells, T-cells, etc.).

Try to run the code line by line and interpret it. There is no need to understand everthing, but try to have an idea of what’s going on and what is stored in the main variables.

By the end of this part you should be able to generate boxplots for any gene in any sample present in the dataset. You can always ask us for help, and to get details about a specific function, you can type in the console: *help(command\_name)* or *?command\_name*.

In Part 2, we will process the dataset to remove duplicated gene names and to remove genes with no expression. We will also extract samples metadata such as donor identifier and cell type from samples' names and use this information to generate new plots.

In part 3, we will perform Principal Components Analysis and generate relevant plots to allow visual interpretation of transcriptomic similarities between samples.

Finally, in part 4, we will perform gene differential expression analysis between cell types using DESeq2 package and export results to Excel.