Homework 1 – Cancer Bioinformatics, Winter 2025-2026
Copy Number Alteration & Expression Levels

DATA: cBioPortal - TCGA PanCancer - Breast invasive carcinoma (BRCA) files. we need the mRNA expression, copy-number-alterations, and the clinical-patient files.

- **data_mrna_seq_v2_rsem.txt**
- **data_log2_cna.txt**
- **data_clinical_patient.txt**

Goal: Find a correlation between copy number alterations VS mRNA expression.

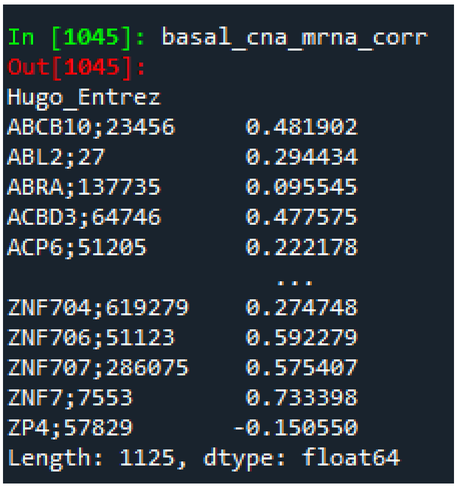
1. Read the 3 files (delimiter is "**\t**"). From now on, instructions will be addressed to mRNA and log2cna data frames.
2. Remove all the NaNs.
3. **(3 points)** Remove genes without expression in all samples.
4. **(5 points)** Make a column called **Hugo_Entrez** that is the result of **Hugo_Symbol** and **Entrez_Gene_Id** separated by a semicolumn (;) .
5. **(10 points)** Remove duplicated **Hugo_Entrez** pairs in each table.
6. **(15 points)** Match gene names by removing rows where the **Hugo_Entrez** is not shared between the two tables. Friendly suggestion- use set operations. This will result in two data frames with equal amount of rows but not columns. (After that, remove the Hugo_Symbol and Entrez_Gene_Id columns)
7. **(2 points)** In both tables set the row names (index) as **Hugo_Entrez**.
8. **(7 points)** Match the sample names between the two tables. This will result in two data frames with the same dimensions.

From now on, instructions will be addressed to the **log2cna** and **clinical_patient** data frames:

9. **(7 points)** Align both data frames to include the same patients. You're going to have to go through some string manipulations, and it's suggested to transpose one of the tables.
10. **(1 point)** Add the SUBTYPE column of the **clin** file to the **log2cna** file.
11. **(3 points)** Group the log2cna file by subtype.
12. **(12 points)** Calculate the median CNA score for each gene in the different molecular subtypes.
13. **(4 points)** Save the data in a table, where each column is a subtype, and each row is a gene. (Should be 16808 rows and 5 columns)
14. **(4 points)** Why do you think some rows have similar/identical values?
15. **(5 points)** List the top 25 genes with the largest number of copies in each molecular subtype.
16. **(12 points)** Create a function that receives a ***gene*** and returns a boxplot showing the CNA scores in the Y axis and the different molecular subtypes in the X axis.

a. **(Bonus 5 points)** Add the Kruskal-Wallis p.value to the plot for all pairs. You can use the **Annotator** class imported from **statannotations.Annotator** module.

17. **(15 points)** For the **Basal** subtype, for genes with a median CNA score greater than 0.4, calculate the Spearman correlation between CNA scores and mRNA expression **per gene**. The result should be stored in a pandas Series with genes as row names and the value of each correlation test as the value. It should look like this:



```
In [1045]: basal_cna_mrna_corr
Out[1045]:
Hugo_Entrez
ABCB10;23456      0.481902
ABL2;27           0.294434
ABRA;137735       0.095545
ACBD3;64746       0.477575
ACP6;51205        0.222178
                    ...
ZNF704;619279     0.274748
ZNF706;51123      0.592279
ZNF707;286075     0.575407
ZNF7;7553         0.733398
ZP4;57829        -0.150550
Length: 1125, dtype: float64
```

18. **(2 points)** Using the function in Q16, plot the two genes with the highest correlation in Q17.

The submission should include one '.**py**'/script file and a '.**PDF**' file for the plots.

*Please add comments to the code

Good Luck!