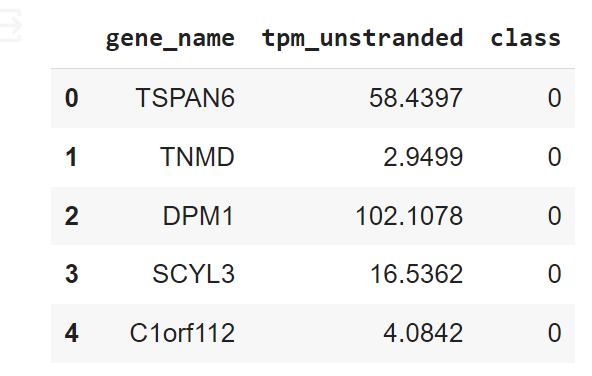
**Lab 9**

**Genetic Algorithm for Bio-Markers Selection** Marks = 30

Bio markers are the genes which are important for a cancer type determination in the human DNA data sample.

A dataset is given. It has three columns.

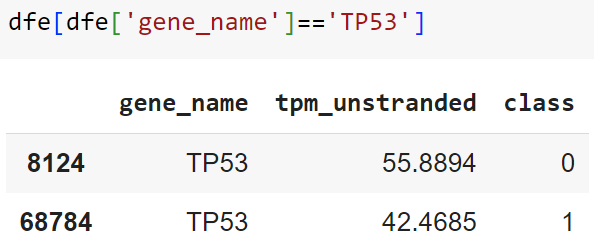


***gene\_name*** is the name of the gene , ***tpm\_unstranded*** stands for transcripts per million and calculated from unstranded RNA-Seq data. ***class*** represents either this genes tpm\_unstranded belongs to normal or cancer tissue. 0 for normal and 1 for cancer.

Problem definition is to train a machine learning algorithm on this provided data. Which is already done in this case. For example, DecisionTreeClassifier. But the accuracy coming up is very low. What we want to do is to Features Selection through *Genetic Algorithm*. You are required to run Genetic Algorithm on the given data and to select top 30 genes set whichmay maximizes the accuracy of our machine learning algorithm. Lets set our goal to 85% or above.

Important points are:

1. Each gene in the dataset is associated with both class=0 and class=1. Only genes with ***tpm\_unstranded*** values present in both class=0 and class=1 will be considered. With 30 selected genes, there will be total of 60 rows, each representing a gene with ***tpm\_unstranded*** values in both class=0 and class=1.



1. Use DecitionTreeClassifier as machine learning algorithm. Fitness function will train the selected 30 genes and accuracy will be treated fitness score.
2. Once the set of genes is selected, print them.
3. Helping code is attached along the lab.
4. This task will be evaluated against absolute **4** marks. Evaluation to be done at ***9-April-2024*** during online – lab session.