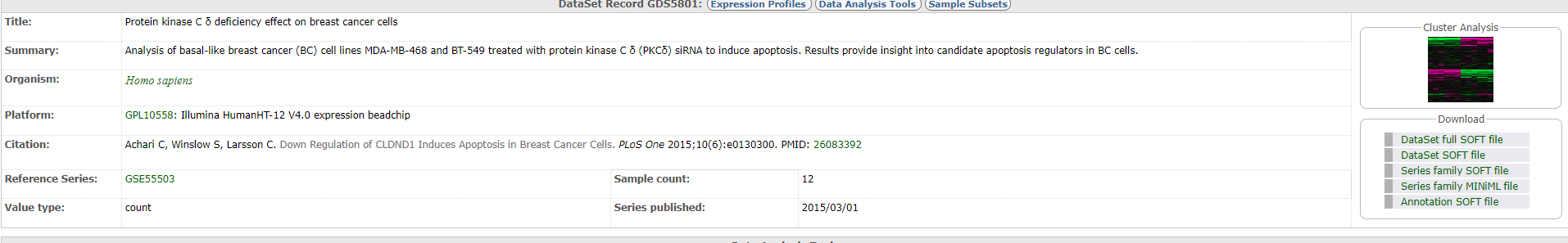
# The problem

The problem of classification is a hard problem for which there is no “best answer” to which method is the best. Specifically, we would like to know which the best classification method for classifying gene expression for the chosen dataset.

# Our Dataset

We chose the following dataset for our project:

<https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS5801&fbclid=IwAR2YoCBZlS1p9Ef7H0dMCqSAj0rsk8ysOuIEyde_3mqnsnb51GnjKHaHNeU>



# Biological Background

To clearly understand the task, we would like to first deepen our knowledge in the following biological topics related to our dataset:

1. Breast Cancer
2. protein kinase C δ
3. Apoptosis

# More details & Plan

12 samples

The cell types:

BT-549 cells treated with siPRKCD

BT-549 cells treated with siControl

MDA-mB-468 cells treated with siPRKCD

MDA-MB-468\_siControl\_rep1; src: MDA-MB-468 cells treated with siControl

* BT-549 is breast cancer cell line (Invasive ductal carcinoma)
* MDA-mB-468 is breast cancer cell line (adenocarcinoma)
* PRKCD - Protein Kinase C Delta, the protein used as treatment

So we have 2 types of breast cancer cells, from each type there are cells treated with PRKCD or control cells (not treated?).

What we’ll do:

Some preparation steps (like in hw4), then we’ll use a few clustering (kmeans, gmm, hierarchical clustering), and a few classifying (svm, naïve bayes, decision tree, knn) methods to identify genes that are affected by the treatment.

Questions:

Is it correct to say that we classify genes, the expression levels on the samples being the features?

How exactly knowing the labels (supervised) help? Since labels are on the samples, not the genes

# Questions

Is this database the best choice for our project?