**Regression**: Can we predict the efficacy of a specific drug given the cancer cell line’s enhancers expression pattern?

**Big Idea**: The enhancer expression usually initiates a pathway in cell development, such as self-replication. Integrating with drug response data, we want to see whether the enhancer expression pattern can predict cell-line’s drug IC50 (The half maximal inhibitory concentration).

**Data Description**: The enhancer expression data can be downloaded in txt format from FANTOM project at http://fantom.gsc.riken.jp/5/datafiles/latest/extra/Enhancers/. After filtering, the data contains a total of 287 observations (cell-lines) and 15,808 features (enhancer expression). There are some missing data. The drug response data (IC50) can be downloaded from GDSC project at https://www.cancerrxgene.org/downloads. There are IC50 data for all the 287 cell line for 251 different drugs.

* **Response**: Numerical drug log(IC50) data for each cell line. Furthermore, since we have 251 different drugs, we can train 251 models for each drug.
* **Features**: 15,808 total, all numerical. The enhancers are annotated with genome coordinates, no previous anticipation.
* **Note**: Since the number of predictor is far larger than the observation. We would use approaches to reduce the dimension such as forward selection or mixed selection. Also, we might also apply shrinkage methods like Ridge and Lasso (Elastic Net) to reduce the feature number further.

**Questions of interest**: The primary question of interest will be to determine how well we can predict a specific drug’s IC50 according to the cell line’s enhancer expression pattern. We can also explore how significant each predictor (enhancer) in explaining the drug response. Finally, we might identify the enhancer which plays an important role in the cancer development path way according to its sensitivity to a certain type of drug.

**Interested Parties**: This might be interested to biomedicine researchers. The relationships between enhancers and drug response can shed light on the molecular pathway of cancer development. What’s more, after verification from bench, the drug with strong relationships with well-known enhancers could be potentially developed as personalized treatment by pharmaceutical companies.

The primary R code for preprocessing

# Part 1: preprocess gene\_expression versus cell\_lines matrix

# downloaded from FANTOM Enhancer project (same as above)

# File server: http://fantom.gsc.riken.jp/5/datafiles/latest/extra/Enhancers/

gene\_expression\_matrix = read.table("human\_permissive\_enhancers\_phase\_1\_and\_2\_expression\_tpm\_matrix.txt")

# remove rows:

# download a preset filter for a well-known cancer cell\_line study:

# Paper URL: https://www.cell.com/cell/fulltext/S0092-8674(18)30307-6?\_returnURL=https%3A%2F%2Flinkinghub.elsevier.com

# Under supplemental information, the filtered cell lines sheet is open for downloading

# The filter pipeline details are descriped in Figure S1.A of the paper

# File URL: https://www.cell.com/cms/10.1016/j.cell.2018.03.027/attachment/ce8d2894-ce79-4334-bda2-e25b625c0f45/mmc1.xlsx

install.packages("openxlsx")

library("openxlsx")

filtered\_gene = read.xlsx("mmc1.xlsx", sheet="15808\_enhancers\_in\_this\_study", startRow = 2, cols = 1)

# relabel the columns (cell\_lines):

# downloaded from FANTOM Enhancer project

# File server: http://fantom.gsc.riken.jp/5/datafiles/latest/extra/Enhancers/

sampleName\_FantomID\_table = read.csv("Human.sample\_name2library\_id.txt", header = FALSE, sep=":", stringsAsFactors = FALSE)

cellline\_CNhs\_pair <- as.character(subset(sampleName\_FantomID\_table, V2!="", select = V2))

names(cellline\_CNhs\_pair) = c("cell line:CNhs")

# I am stuck here since the strsplit() does not seem to work.

celllines <- strsplit(cellline\_CNhs\_pair, '\t')

# Part 2: preprocess drug\_response (log(IC50)) versus cell\_line matrix

# downloaded the spreadsheet from https://www.cancerrxgene.org/downloads at "log(IC50) and AUC values"

drug\_IC50\_cellline = read.xlsx("v17.3\_fitted\_dose\_response.xlsx", sheet="Sheet1", cols = c(4,6,10))

levels(as.factor(drug\_IC50\_cellline$DRUG\_NAME))

# transform the spreadsheet to a matrix

# more work to do ...