Homework 4

Abhijit Dasgupta

In this homework we will play with some publicly available RNA-Seq data. Information about the experiment is available [here](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE123519). Briefly, this is a mouse experiment where the authors were interested in PIK3CA mutation variants. This interest is derived from PIK3CA’s implication in glioblastoma tumorogenesis. There are 26 mice with different engineered genetics.

The data and experimental design are available as a zip file and excel file respectively in Teams Week 4 files (GSE123519\_RAW.zip and GSE123519design.xlsx).

### Prelim work

1. Download and unzip the data file in your project data directory. You should get 26 text files, perhaps in a sub-directory. Note this, since you’ll need it in the next step. Also download the experimental design file.

### Homework

#### Data preparation

1. Read all 26 files into a list in R, where the list consists of 26 data frames
2. For each of the data frames perform the following data processing
   1. Keep only the variables tracking\_id, gene\_id, and locus through FPKM.
   2. Change the gene names in gene\_id to capital letters (use stringr::str\_to\_upper or other equivalent functions)
   3. Split the locus variable into a chromosome and a location variable.
   4. Filter **out** all observations with FPKM less than 1 (no real scientific reason, just because)
3. Now make a single data set stacking the processed datasets, adding a column that specifies the sample identifier of each observation. You might like to call this column sampleID to make the next step simpler
4. Import the design information into a data frame. Add the design information to the genetic dataset you created above, ensuring that ids and genotype specification (stored in the column Class) are aligned.
5. Note that there is an add-on to the genotype information, in the form -1, -2 to specify which particular sample it is. This is unnecessary and would create problems for any comparative analyses using t-tests or ANOVA later. Please remove it so that you just have genotype information.
6. There is one idiosyncratic genotype specification which is a re-rerun. Fix this to match the other genotypes.
7. Clean the chromosome values so that you only have data of the form “chr##” where ## refers to the chromosome number.
8. Save the dataset using saveRDS to a .rds file

You can choose whether you fix all the genotype information before or after you join the design dataset to the genetic dataset. Whatever makes sense to you.

#### Data analysis

1. Provide the following tables:
   * median FPKM levels by Class
   * How many unique genes are interrogated per chromosome
2. How many unique genes of the form PIK3C\* are present in the dataset
3. What are the relative frequencies of each PIK3CA variant among the 26 mice. Assume a missing value means the variant is not present.
4. Perform a statistical test to see if FPKM expression levels are different at PIK3C\* genes compared to all other genes. (the correct test is actually complicated, but doing t-test, wilcoxon test or permutation tests are all allowed here)
5. Draw a barplot showing the median FPKM levels by Class