BSRP Problem Set 4

your-name-here

Introduction

The goal of this problem set is to use correlation analysis on gene expression to examine the KRAS pathway. Then, you will revisit and expand on the KRAS T-test to test for many genes using differential gene expression.

Before you try knitting this document, please install the following packages via your R console (the code is commented out). This installs the needed tools for differential gene expression. You only need to do this once.

Here is our new biological question

Biological Question <-> Computational Analysis

Biological Question	Computational analysis
How is the expression of KRAS correlated with genes related in its pathway in lung, pancreas, and colon cancer cell lines?	Multiple correlation to measure linear relationship.

Assignment Formatting

Write your code in the designated code chunks provided, and write your written answers in the text region of the document. If you want to incorporate both text and code output, the function `cat()` will do the trick in code chunks. For example, `cat("The number of rows in the iris dataframe is", nrow(iris))`.

```
## Fetching https://cds.team/taiga/api/dataset/public-21q1-4b39/33
## Status 200
## loading cached data version from  /home/chris/.taiga/public-21q1-4b39_33.toc

## Fetching https://cds.team/taiga/api/dataset/public-21q1-4b39/33
## Status 200
## loading cached data version from  /home/chris/.taiga/public-21q1-4b39_33.toc

## Fetching https://cds.team/taiga/api/dataset/public-21q1-4b39/33
## Status 200
## loading cached data version from  /home/chris/.taiga/public-21q1-4b39_33.toc

## Fetching https://cds.team/taiga/api/dataset/public-21q1-4b39/33
## Status 200
## loading cached data version from  /home/chris/.taiga/public-21q1-4b39_33.toc
```

Correlation Analysis

Three well-established genes that are activated by KRAS are PI3K, RAF, and RAL. We will use correlation analysis to see how the expression of KRAS, PI3K, RAF, and RAL are related linearly.

To look at the correlation between genes across cell lines, we need the following dataframe:

- Observations: Cell lines.
- Variables: Gene expression of KRAS, PI3K, RAF, and RAL. They are labeled as `KRAS`, `PIK3CA`, `RAF1`, and `RALA` respectively in the column names of `expression` dataframe. Also, we need `primary_disease` specifying what cancer subtype the cell line is.

Probelm 1

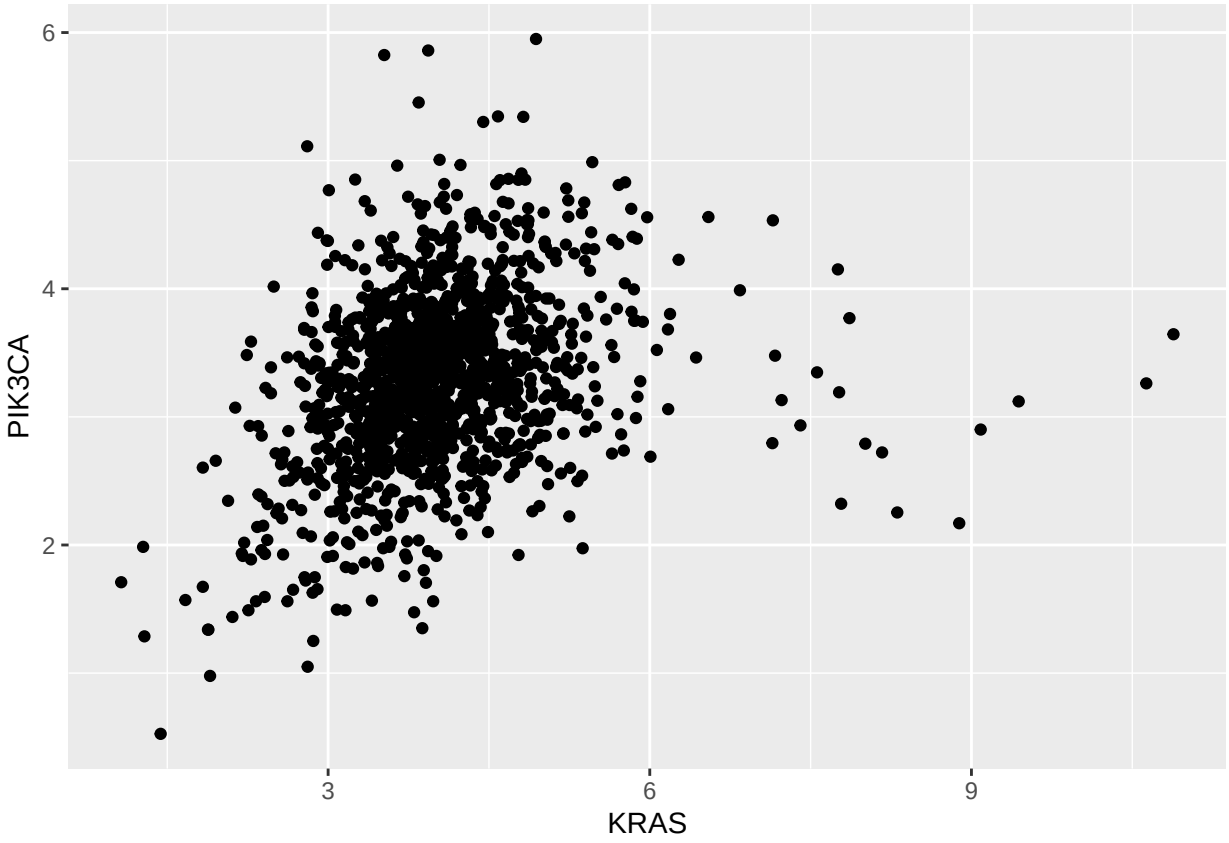
Wrangle a dataframe into the desired dataframe as described above.

Probelm 2

Using the `cor()` function, compute the correlation between KRAS and PIK3CA, and make a scatterplot of the relationship. If you have `NA` in the dataframe and they should be ignored when computing the correlation, use the argument `use = "complete.obs"` in the `cor()` function. How does it look?

```
## [1] 0.2982828
```

```
## Warning: Removed 436 rows containing missing values (geom_point).
```



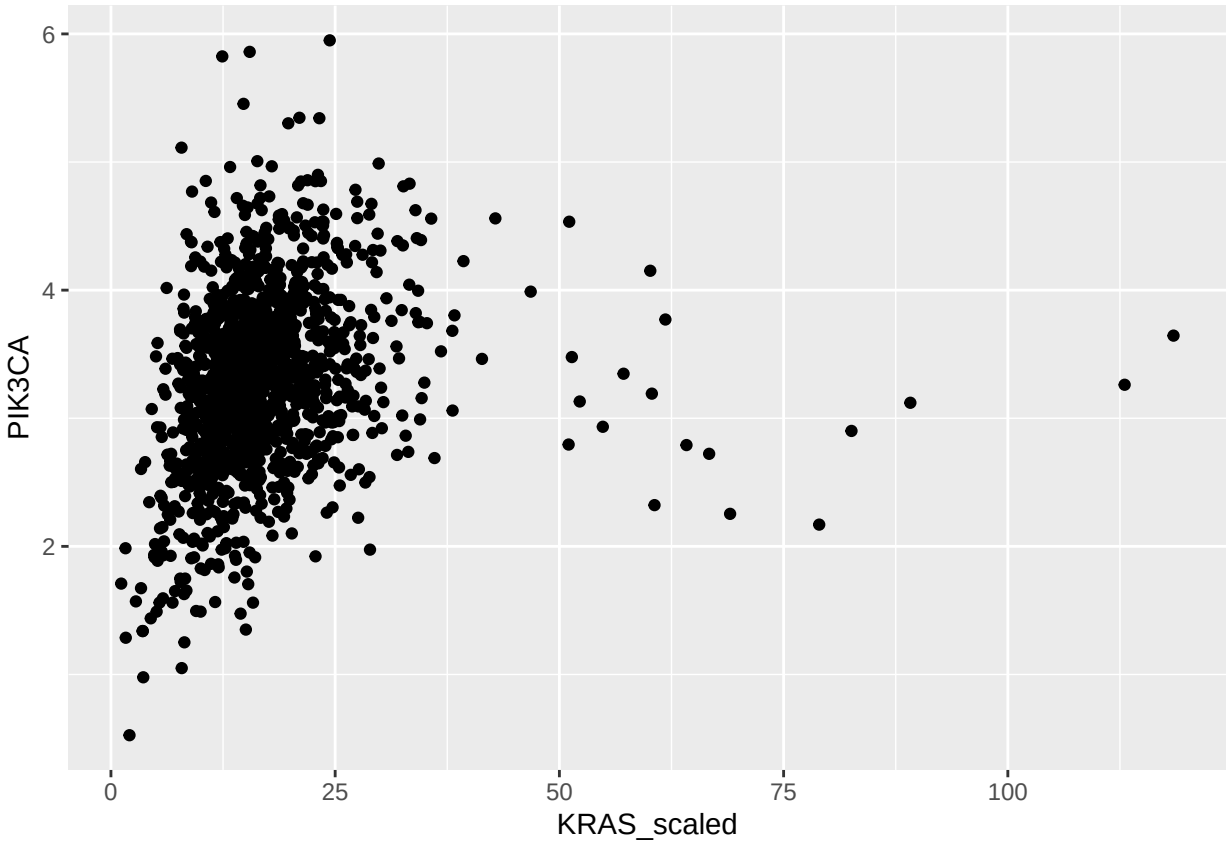
Solution: The linear trend is definitely there, but there are some outliers that may come from not a linear trend. Correlation as a metric isn't useful to deal with data not from a linear trend.

Problem 3

To show that correlation as a metric does not depend on the scale, convert the scale of KRAS from log2 fold change to fold change in a new variable in the dataframe. Then, compute the correlation between KRAS fold change and PIK3CA log2 fold change, and make a scatterplot. What do you see?

```
## [1] 0.2185488
```

```
## Warning: Removed 436 rows containing missing values (geom_point).
```



Solution: The correlation stays the same, but the scatterplot axis changed.

Problem 4

To examine all pair-wise correlations between these four genes systematically, one can summarize it in a correlation matrix. Each row and column correspond to each of the four genes, and the cell correspond to the correlation value.

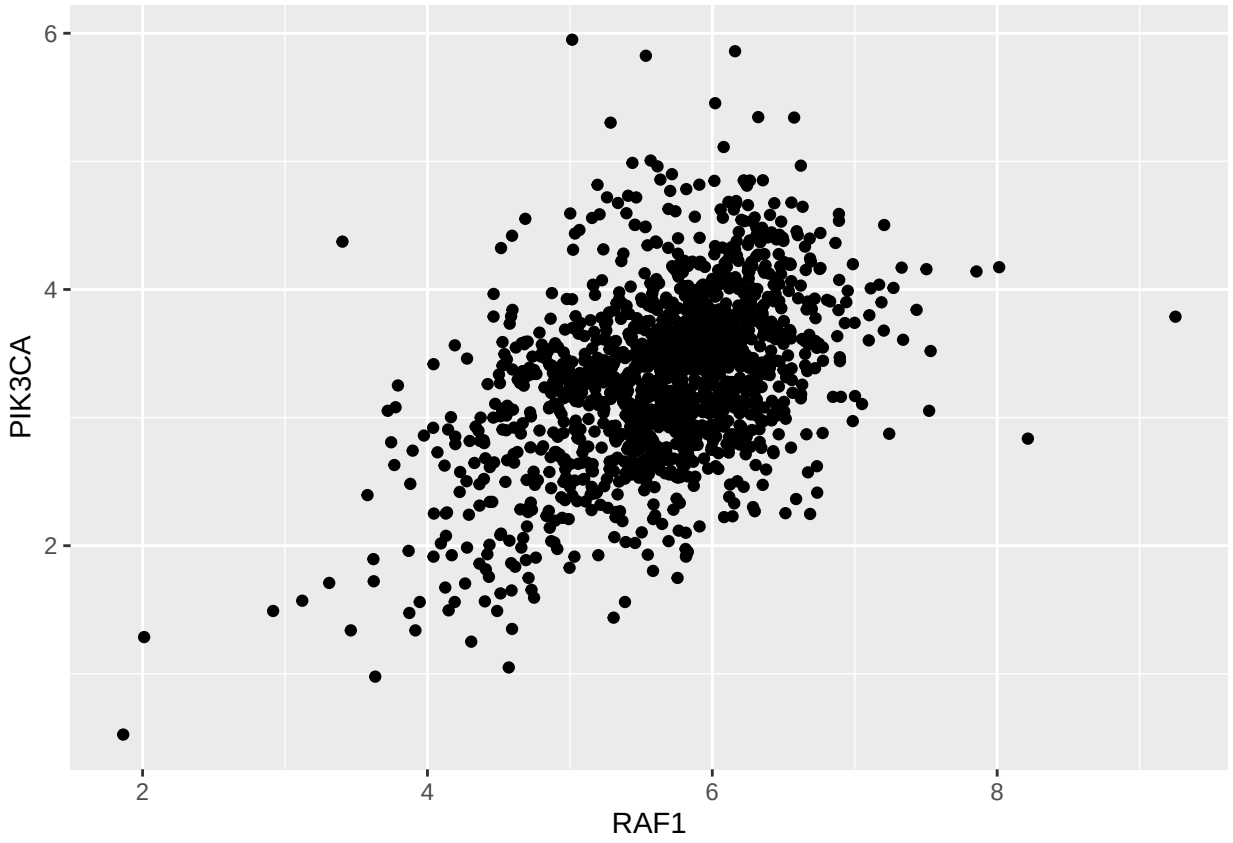
To generate such a matrix, the `cor(dataframe)` function can take in a dataframes with numeric variables. Correlation is computed between all pairwise combinations of the variables in the dataframe.

Wrangle the dataset: you will need to get rid of unnecessary variables.

Then, compute the correlation matrix. What do you see? How would you interpret it? What gene pair relationship is the strongest (that is not correlation with itself)? Feel feel to create a scatterplot of a relationship you find interesting.

```
##             KRAS    PIK3CA       RAF1       RALA
## KRAS   1.0000000 0.2982828 0.30272193 0.18154645
## PIK3CA 0.2982828 1.0000000 0.45703481 0.21974103
## RAF1   0.3027219 0.4570348 1.00000000 0.07640372
## RALA   0.1815465 0.2197410 0.07640372 1.00000000
```

```
## Warning: Removed 436 rows containing missing values (geom_point).
```



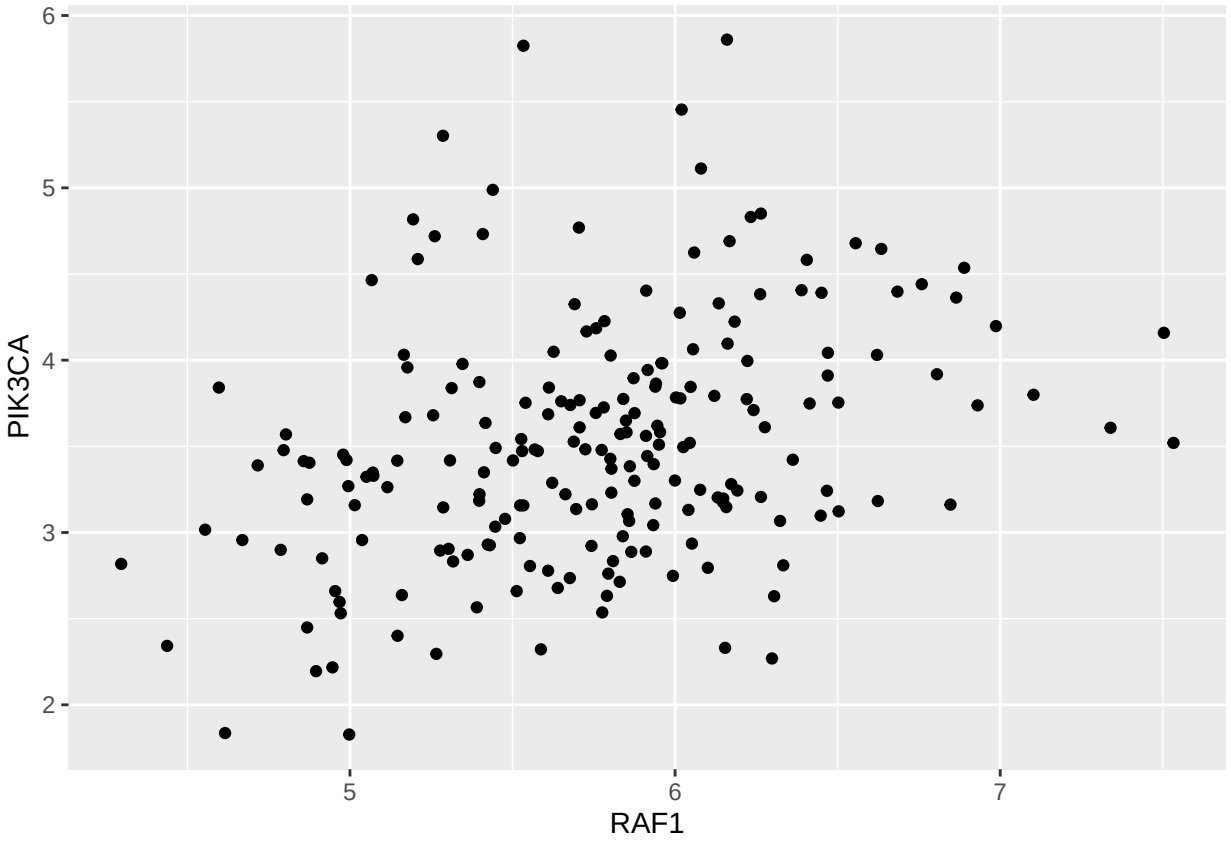
Solution: RAF1 and PIK3CA has the strongest correlation. KRAS may be upregulating these two genes in similar ways, or other pathways are activating these two genes in very similar ways.

Problem 5

Subset the data down to lung, pancreas, or colon cell lines, and redo the correlation matrix analysis. How does an analysis specific to a cancer subtype compare the all cell lines?

```
##             KRAS    PIK3CA      RAF1      RALA
## KRAS   1.0000000 0.1798130 0.1620890 0.1130942
## PIK3CA 0.1798130 1.0000000 0.3336860 0.3662232
## RAF1   0.1620890 0.3336860 1.0000000 0.2196228
## RALA   0.1130942 0.3662232 0.2196228 1.0000000
```

```
## Warning: Removed 67 rows containing missing values (geom_point).
```



Solution: Similar trend for lung cancer cell lines, but definitely less strong of a linear association.

Differential Gene Expression

Now, we are going to compare KRAS WT/MT cell lines against gene expression of all genes profiled from RNA sequencing via differential gene expression.

Biological Problem <-> Computational Analysis

Biological Question	Computational analysis
Which genes are the most upregulated or downregulated from KRAS mutation compared to WT in lung cancer cell lines?	Differential gene expression analysis, visualized with volcano plots.

Data Wrangling

We first need to get a dataframe ready to perform differential gene expression. This dataframe gives information which cell lines will be labeled as mutant vs. wildtype when comparing KRAS expression levels. We will look at the gene expression data shortly.

Problem 6

We need the following:

- Observation: Lung cancer cell lines.
- Variables: DepMap_ID, KRAS_mutated (logical vector whether KRAS is mutated).

You should be able to use code from Week 3 to create this dataframe fairly easily.

Setup code for differential gene expression

The following code first loads in the raw RNA sequencing data, which is the raw data that is used to generate the gene expression data. Differential gene expression analysis takes this raw RNA sequencing data, processes it differently, and runs the desires analysis across all genes.

The code then defines a function to perform differential gene expression analysis in a function, `runDifferentialGeneExp`.

The first argument, `DepMap_ID`, is a String vector of the cell line names to perform differential gene expression.

The second argument, `cellLineStatus`, is a Logical vector indicating whether the corresponding cell lines in DepMap_ID is in the mutated group, with TRUE being mutated, and FALSE being wildtype.

The third argument, `filterCutOff` is optional. It gives a cutoff for filtering the RNA sequencing data of genes with little to no read counts.

When `runDifferentialGeneExp` is called, it subsets the RNA sequencing data by `DepMap_ID` argument, and then performs a series of data transformation steps, as well as filtering out genes that do not pass the cutoff. Two plots are created: before data transformation, and after data transformation. An ideal plot after data transformation has the blue line to be horizontal. (Please let me know if it doesn't look like that!) Then, it runs a statistical test similar to the T-test over and over again across all genes comparing WT/MT cell lines using `cellLineStatus` argument, and returns a dataframe containing the top deferentially genes.

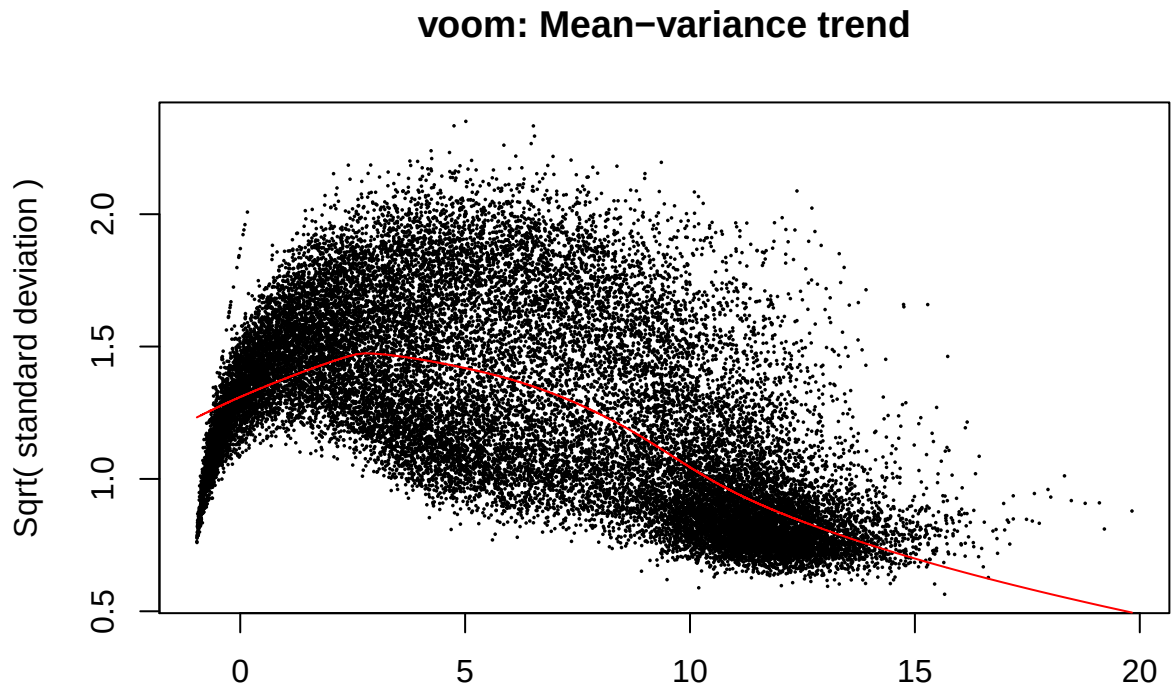
```
## Fetching https://cds.team/taiga/api/dataset/public-21q1-4b39/28
## Status 200
## loading cached data version from  /home/chris/.taiga/public-21q1-4b39_28.toc
```

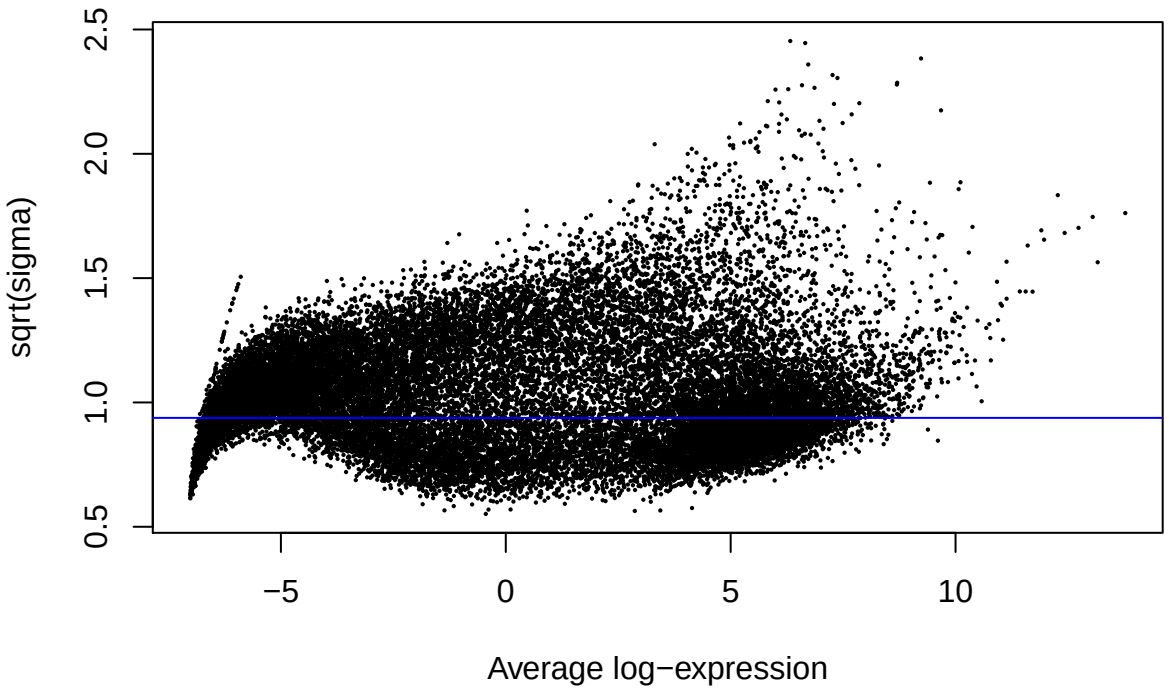
Problem 7

Run `runDifferentialGeneExp` on lung cancer cell line samples to test differential gene expression by KRAS MT/WT. Does the blue line look horizontal after data transformation?

Then, with the returned dataframe, describe a few of the top differentially expressed genes, in terms of the biology, (adjusted) p value, and log fold change:

For instance, are these genes expressed higher or lower between MT/WT cell lines? Search the gene symbols, and try to understand a little bit of biology of a gene or two - do you think it makes sense?



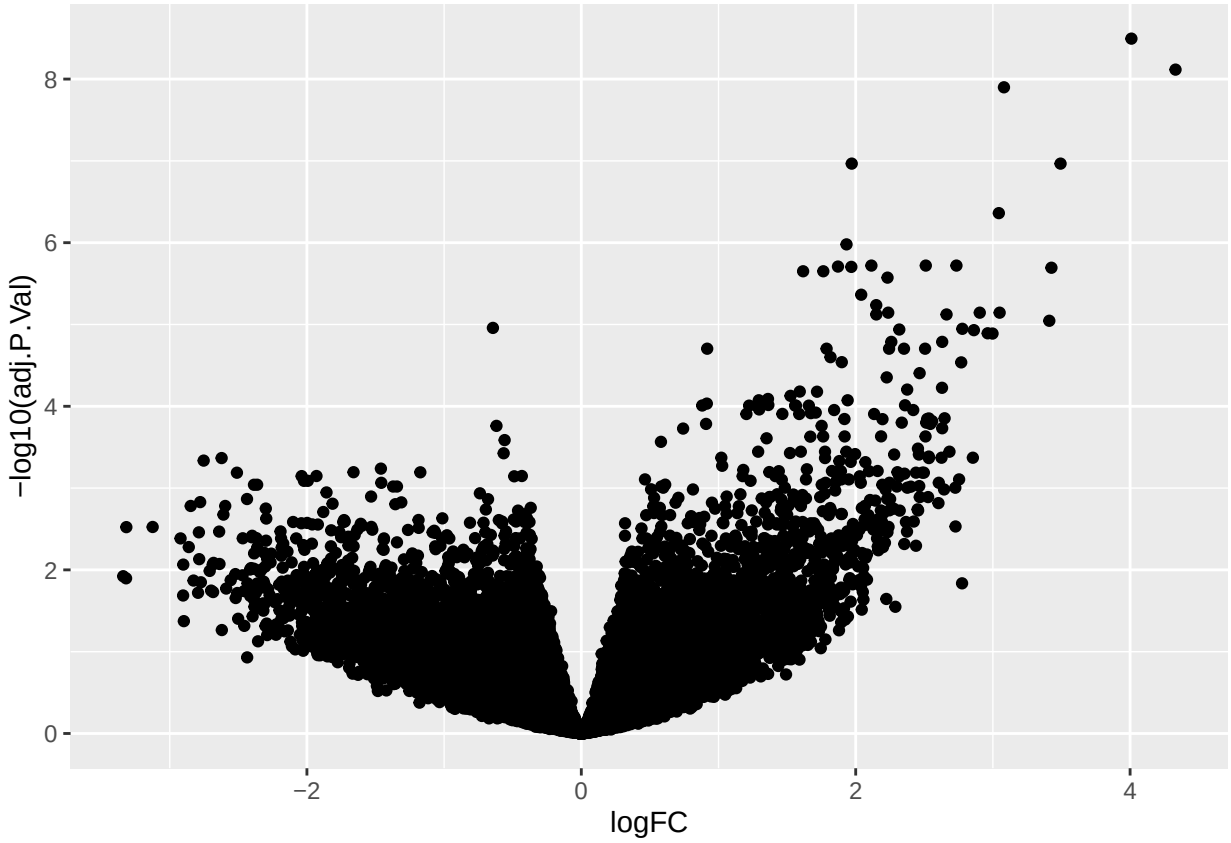


	logFC	AveExpr	t	P.Value	adj.P.Val	B
AREG	4.009560	2.070870	7.948696	0	0e+00	20.49902
EREG	4.331083	1.195586	7.693699	0	0e+00	18.94694
AC010198.2	3.080625	1.428522	7.544450	0	0e+00	18.05151
MYEOV	3.493092	1.488424	7.130186	0	1e-07	15.74297

	logFC	AveExpr	t	P.Value	adj.P.Val	B
SPRY4	1.971121	3.781124	7.095445	0	1e-07	15.58787
MLPH	3.043398	3.505034	6.821592	0	4e-07	14.06622

Solution: The top differentially expressed genes are upregulated by MT KRAS, with a log fold change of 4. The genes AREG and EREG are ligands that bind to the epidermal growth factor, which activates the KRAS pathway. KRAS mutation is associated with most increased level of these ligands expression, but the mechanism of action is unclear.

Then, using ggplot, make a volcano plot of the results. Transform the adjust p-value into -log10 scale using the function `-log10()`. What do you notice about the overall trend of the volcano plot?



Solution: The significant genes are generally upregulated than downregulated.

Lastly, where does KRAS expression fall in this ranked list of differentially expressed genes?

Problem 8

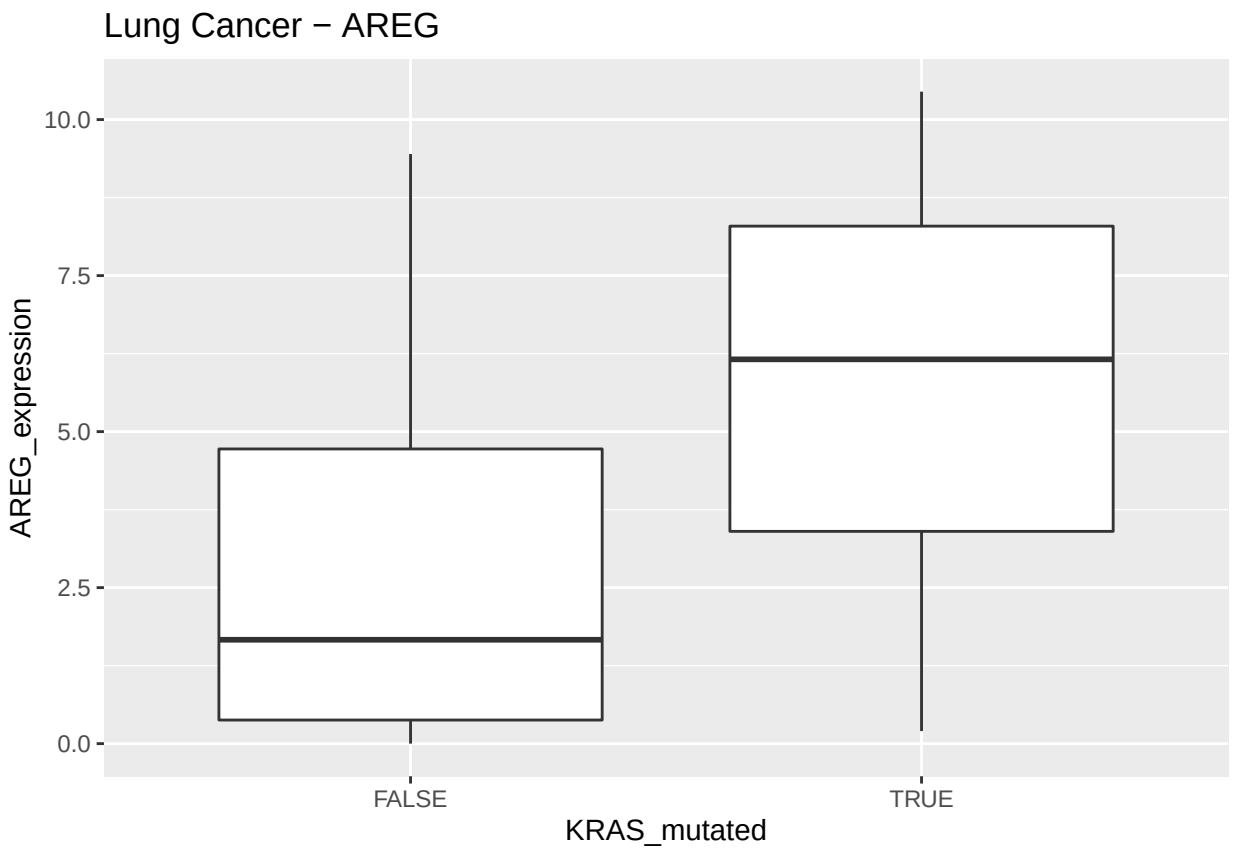
To make sure that the results are correct, pick a gene of interest from the list of top differentially expressed genes. Create a boxplot of the gene' expression based on KRAS WT/MT in lung cancer cell lines. What do you notice between this analysis compared to differential gene expression analysis?

You will need to use the `expression` dataframe, and not `RNA_reads` dataframe.

```
## Sample size of MT 51
```

```
## Sample size of WT 222
```

```
## Warning: Removed 67 rows containing non-finite values (stat_boxplot).
```




```
##
##  Welch Two Sample t-test
##
## data:  AREG_expression by KRAS_mutated
## t = -6.285, df = 72.841, p-value = 2.153e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.907456 -2.025892
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            2.719834            5.686508
```

Solution: The log-fold change is less in this analysis than the reported from differential gene expression. This may be due to different processing strategies.