**Statistical Analysis of Gene Expression Data in Lung Squamous Cell Carcinoma and Clear Cell Renal Carcinoma**

**Prepared by:**

1. **Mohamed El-Menawy(Intro and data prep)**
2. **Marwa Abdel-gelil (Linear regression and feature selection)**
3. **Rawan Yasser (Volcano plotting and GSEA)**
4. **Gehad Kamel (Hypothesis testing paired and independent)**
5. **Introduction**

Cancer is a deadly disease that occurs when specific cells in the body starts do divide uncontrollably and spread to other parts of the body. There different types of Tumors; Benign and Malignant tumors. Mainly benign tumors are easily managed and treated; however, we tend to focus more on malignant tumors that are the main cause of deaths and tend to metastasis.

Cancer is mainly caused by DNA alterations or chromosomal aberrations that cause some genes to be over-expressed or suppressed. There are three types of genes that cause cancer which are: proto-oncogenes, tumor suppressor genes, and DNA repair genes.

In this study we are performing statistical analysis on the gene expression data obtained in two types of widely spread and deadly cancer types:

**A-) LUSC (Lung squamous cell cancer)**

Lung cancer is the most common diagnosed malignant tumor, and is the leading cause of cancer-associated mortality. LUSC is one of common sub-types of non-small cell lung carcinoma (NSCLC).

**B-) KIRC (clear cell renal cell carcinoma)**

Clear cell renal cell carcinoma is the common type of kidney cancer that accounts for 80% of all renal cell carcinoma cases. Clear cell carcinoma is named after how the tumor looks like under the microscope since the cells are clear, like bubbles.

In this study we obtained RNA-Seq normalized data for the lung squamous cell carcinoma (LUSC) and clear cell renal cell carcinoma (KIRC) from the Cancer Genome Atlas (TCGA) for the Gene expression analysis in healthy tissues and cancerous tissues to identify differentially expressed genes; Also, we obtained copy number variants (CNV) data for each cancer type to identify if there is correlation between differentially expressed genes and their relation to the cancer development and growth. We performed Gene expression analysis using three different methods: Hypothesis testing, Fold change, and performing Volcano plotting to showcase the difference between the two methods.

Identifying the top five differentially expressed genes in each cancer type can help us perform correlation and linear regression analysis to detect whether there is a relation between chromosomal aberrations and gene expression and their correlation to Lung squamous cell carcinoma (LUSC) and clear cell renal cell carcinoma (LUSC).

1. **Methods**

Statistical analysis was conducted by R version 4.1.3, using packages such as: a) stringr b) bioconductor “enhanced volcano” c)glmnet d)car.

Workflow is described in Fig. 1 Work flow diagram.

Reading Data

Data Preparation & Categorization

Paired Hypothesis testing

Lowest 5 P.Value selection

GSEA

Error Correction & P.Value sort

Independent Hypothesis testing

Log 2 fold change & Ensembl ID generation

Multiple regression

Fig. 1 Work flow diagram

**II.I Data preparation**

Firstly, data retrieved and uploaded on R studio. “Data preparation.R” containing “prepare\_GE\_data” function **(prepare\_GE\_data <- function(path.healthy, path.cancer,path.CNV)**

That takes the path of the Gene expression of healthy samples, GE of cancer samples, and path for the copy number variant (CNV) file. “prepare\_GE\_data” function works as follows: A-) read the GE and CNV data and then clean the data from genes with zero expression. B-) Getting gene names after filtration and finding the intersection between healthy and cancer data. C-) Getting common samples between CNV data and Cancer data. E-) returns a list of cleaned Gene expression data of Healthy and Cancer samples, CNV data, and Gene expression cancer data specifically used for linear regression.

**II.II Hypothesis Testing**

Hypothesis testing is performed to identify Differentially Expressed Genes (DEGs) for both healthy and cancer samples for the two different types of cancers : a-) LUSC and b-) KIRC.

Firstly, we performed paired hypothesis testing using the function: **t\_test <- function (data.healthy, data.cancer)** in a source file (“Paired\_test.R”) ; such function perform paired hypothesis testing between healthy and cancer samples by 1-) Calculating paired difference between the data points 2-) performing Shapiro testing to test whether the difference follows normal distribution or not. 3-) If the difference follows normal distribution (P > 0.05) we perform t-testing and then obtain the p-value; else (P <0.05) we perform Wilcox test and capture the p-value in a variable called “paired\_p\_value”. 4-) Such function returns a data frame containing Gene names with their corresponding adjusted p-values resulted from the paired testing.

Secondly we perform independent hypothesis testing using this function: **indep\_test <- function(data.healthy, data.cancer)**

Indep\_test function takes the healthy data and cancer data from the Data prep function and perform independent hypothesis testing by 1-) performing Shapiro test to test if both data points follow normal distribution 2-) if both data follows normal distribution we check the variances of both data if they are equal we perform t.test else we perform welch test 3-) if one of the data points do not follow normal distribution we perform Wilcox rank test. 4-) we capture the resulting p-values from each test in a variable “indep\_test\_p\_value” 5-) such function returns a data frame containing Gene Names and their corresponding adjusted P-values using “fdr” method.

**II.III Fold change**

The second method to infer differentially expressed genes(DEGs) is by calculating fold change between the GE of healthy and cancer data points. We performed Fold change using the function: **fold\_change <- function(healthy.data, cancer.data)**

Such function takes healthy and cancer data of both types of cancer (LUSC and KIRC) and calculates the fold change between each gene expression values; and returns three variables containing DEGs and non-DEGs, and total fold change values for each gene to then be compared by DEGs obtained from hypothesis testing using volcano plotting and GSEA

**I.IV Volcano plotting and GSEA**

**I.V Linear regression and feature selection**

Linear regression and feature selection is performed by comparing the copy number variants (predictor) of each type of cancer to the gene expression of the five top differentially expressed genes (Response) using the function: **run\_reg <- function(top\_5\_DEGs , data.cnv)**

Run\_reg function takes the top five expressed genes obtained from the paired testing and then perform linear regression analysis between the intersect of samples between CNV data and Cancer data. Feature selection is performed if the number of copy number variants (Variables) exceeds the number of provided samples (Response) using the function: **feature\_selection <- function (Data.CNV, DEGs\_1)**

Such function performs cross validation and obtains the minimum lambda value that would be referenced to feature selection and then used back in the “run\_reg” function for completing regression analysis.

1. **Results And Discussion**

**III.I Hypothesis testing**

A-) Paired testing:

The differentially expressed genes produced from the paired testing showed 13176 genes differentially expressed for the Lung squamous cell carcinoma (LUSC) and the GE breakdown is shown in [Fig.2](#fig2). While 13050 genes differentially expressed for kidney cancer and GE break down shown in [Fig. 3](#fig3)

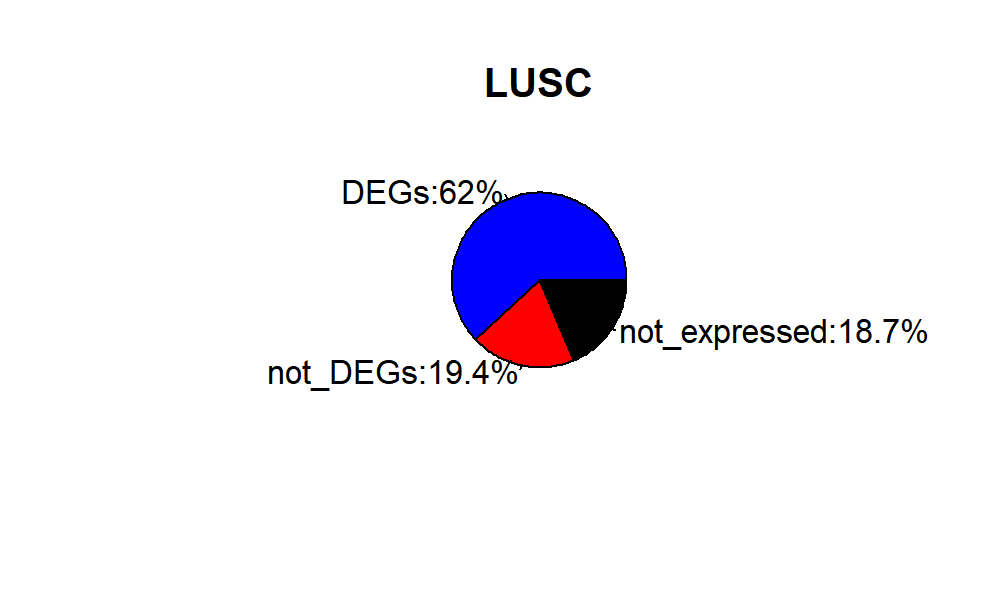
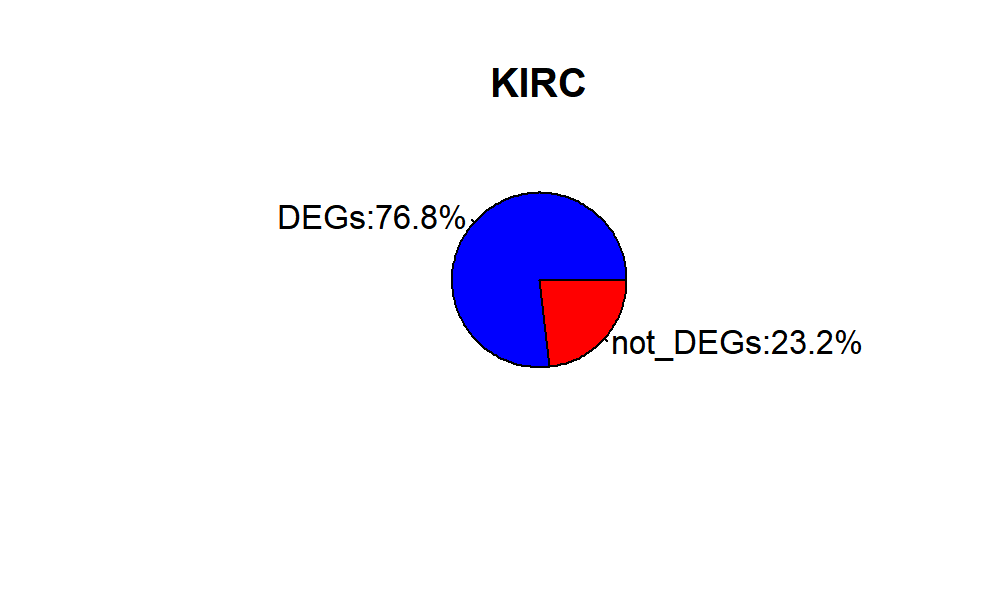
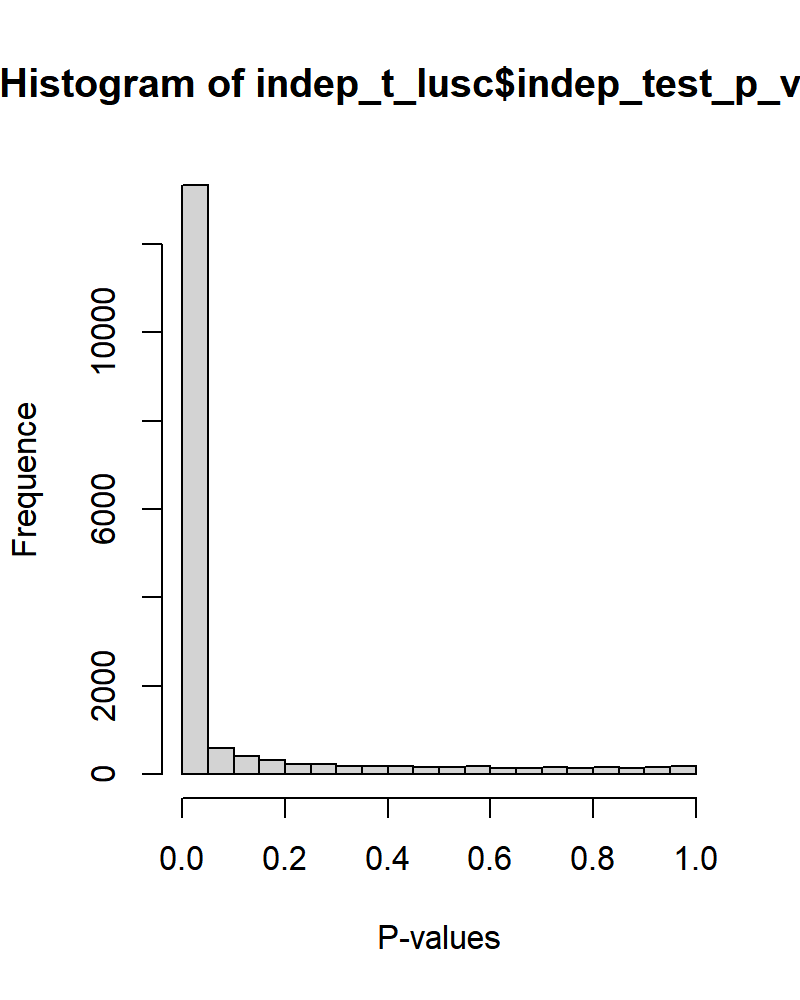
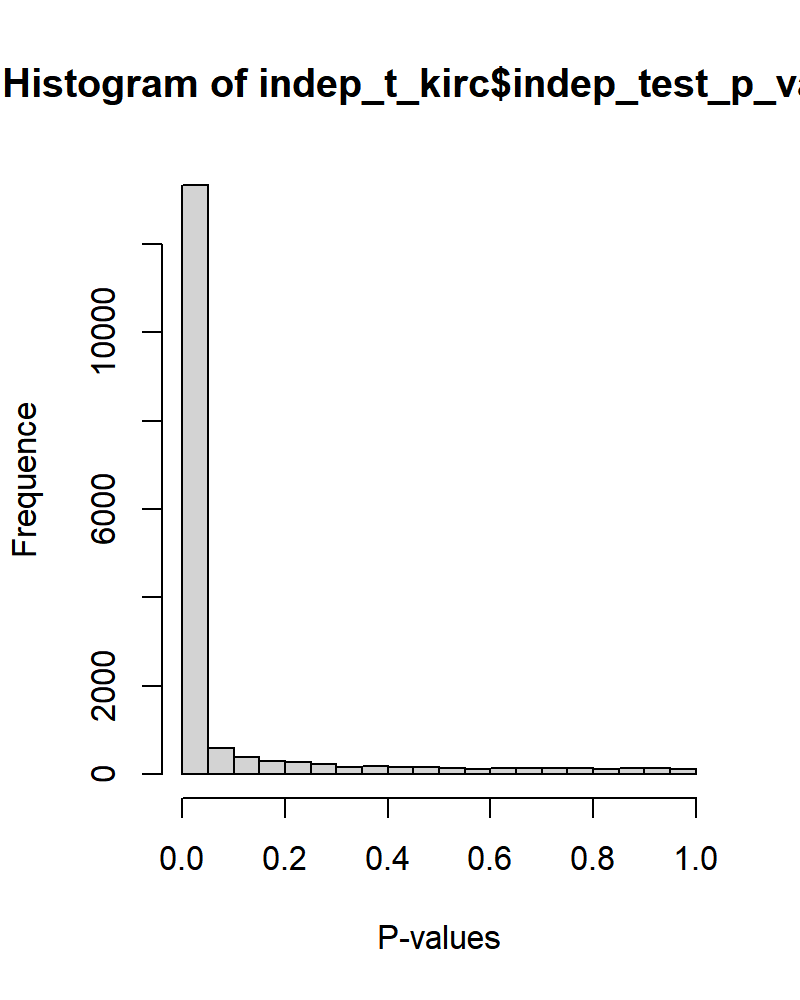


Fig. 3 Pie chart showing percentage of DEGs and non-DEGs in Kidney cacner

Fig. 2 Pie chart showing percentage of DEGs , non-DEGs and not-expressed genes

B-) Independent testing:

The differentially expressed genes produced from the independent testing showed 13323 differentially expressed genes for Lung cancer [Fig.5](#fig5)  and 13331 differentially expressed genes for kidney cancer [Fig.4](#fig4)

**III.II Fold change**

The second method using fold change showed 9186 differentially expressed genes in Lung cancer (LUSC) [Fig. 6](#fig6) and 7318 differentially expressed genes in kidney cancer [Fig. 7](#fig7)

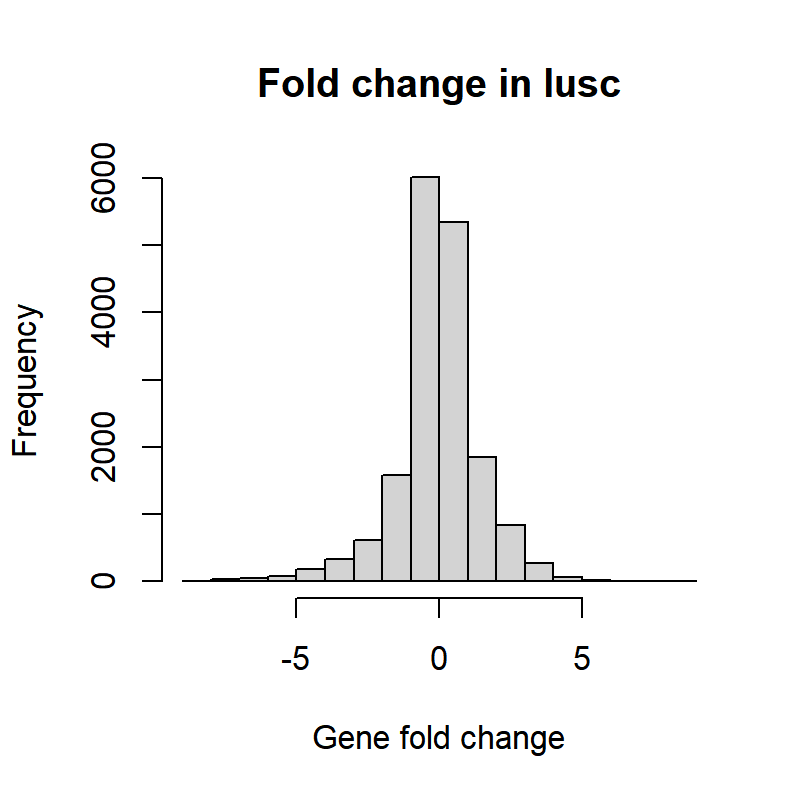
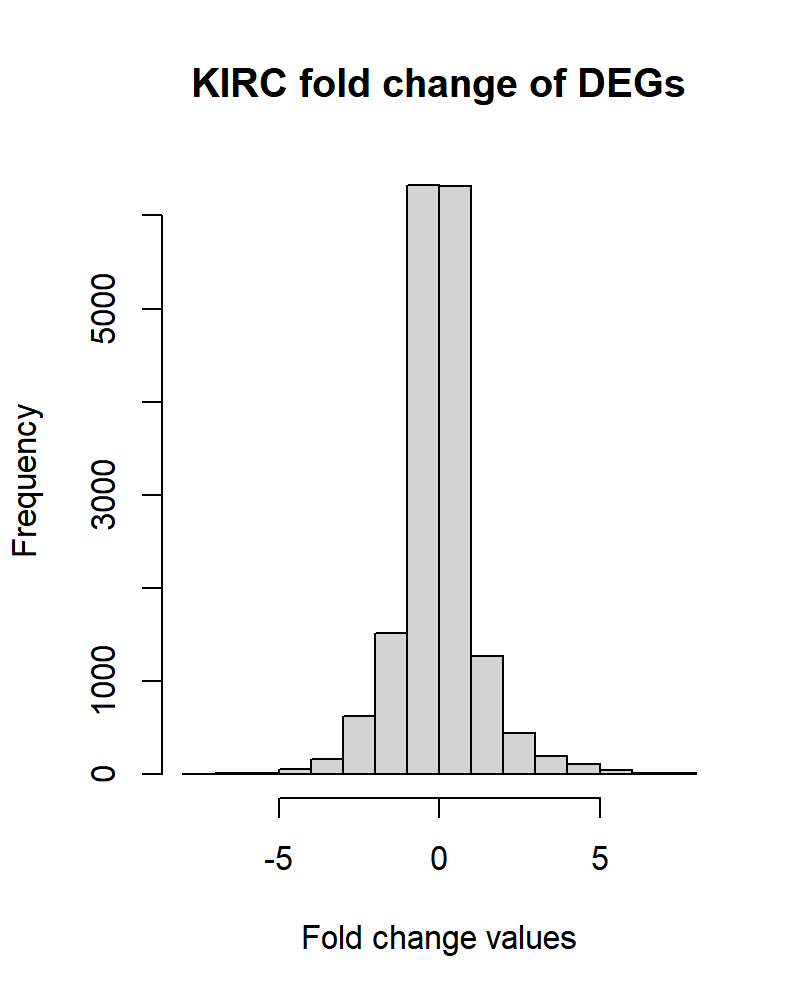


Fig. 7 Fold change data for KIRC

Fig. 6 Fold change data for LUSC

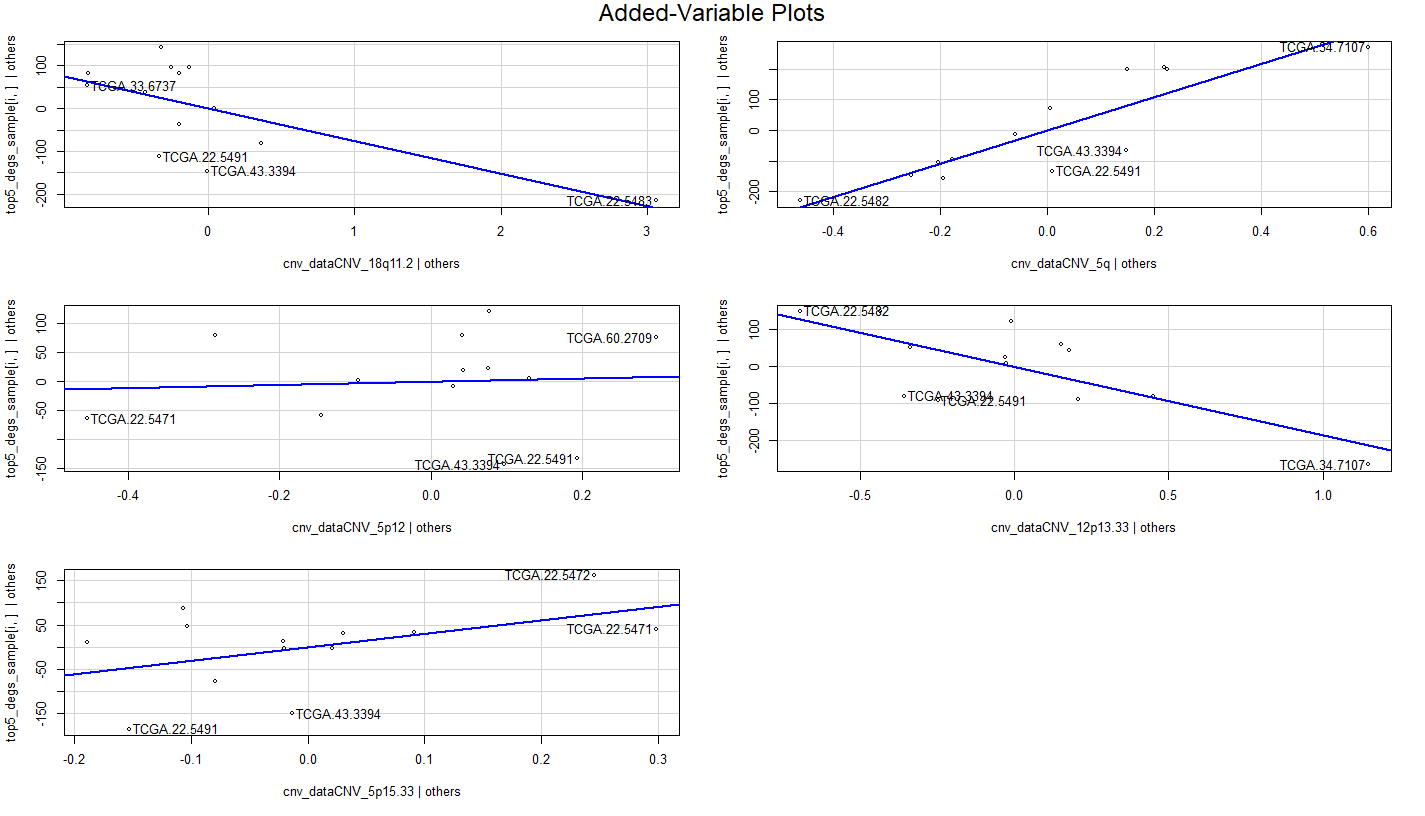
**III.III Volcano plotting and GSEA**

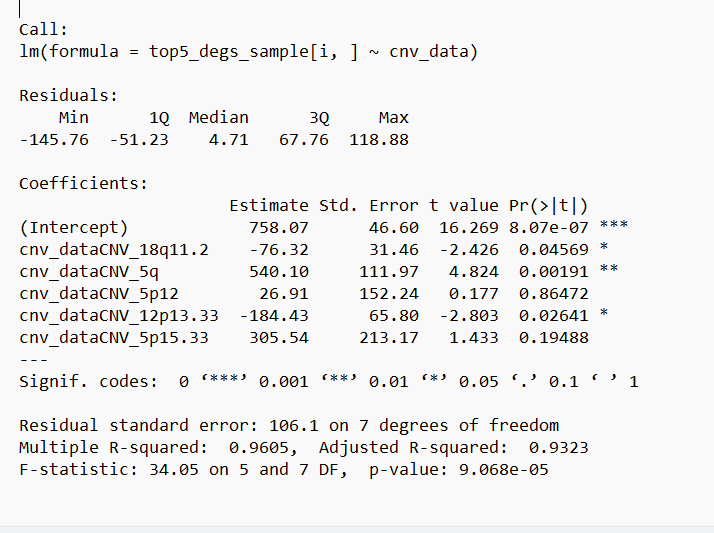
1. **IV Multivariant regression**

3.7. Linear Regression Results:

The linear regression models are generated using R code. Result is shown in fig (x) & table (y) for lung squamous cell carcinoma using feature selection. 1 model of DEGs showed a significant dependence on copy number variation with R2 = 0.9605, while 3 models of DEGs showed only the intercept as a significant output and one model of DEGs has no significance. the alpha value is equals to 0.05.

***Figure x: linear regression models for the DEGs of the least p-value.***





|  |  |
| --- | --- |
| Table ( ) : Lung squamous cancer cell carcinoma regression models : (α =0.05) | |
| GPR116 | Y = 158.69 |
| CCDC69 | Y = 275.82 |
| HYAL1 | Y = 146.53 |
| NDNF | Y = 0 |
| CAT | Y = -76.32 cnv \_18q11.2 + 540.1 cnv\_5q -184.43 cnv\_12p13.33 + 758.07 |

The linear regression models are generated using R code. Result is shown in fig (x) & table (y) for kidney squamous cell carcinoma using feature selection. 3 model of DEGs showed a significant dependence on copy number variation, while 2 models of DEGs showed only the intercept as a significant output. the alpha value is equals to 0.05.

|  |  |
| --- | --- |
| Table ( ) : Lung squamous cancer cell carcinoma regression models Kirc: (α =0.05) | |
| GLRX5 | Y = 728.64 |
| KCNH3 | Y = 4.5309 |
| KCNJ16 | Y = 740.7 x + 1923.8 |
| DCAF11 | Y = 375.15 cnv\_kirc\_dataCNV\_1p31.1 + 985.04 |
| SOWAHA | Y = -40.501 x + 15.790 |

table (y)

1. **Conclusion**

Kidney renal clear cell carcinoma and lung squamous cell carcinoma are two of the most prevalent cancer types. Statistical analysis of current raw tumor and healthy data has pointed out several genes as tumor-associated biomarkers for KIRC patients.

This study provides insights into statistical hypothesis testing, GSEA and regression analysis for paired and independent KIRC and LUSC patient data and could be useful in determining which genes and gene sets that are integrated with tumor progression and diagnosis.

**V.References**

1. <https://www.cancer.gov/about-cancer/understanding/what-is->
2. Chen, J.W., Dhahbi, J. Lung adenocarcinoma and lung squamous cell carcinoma cancer classification, biomarker identification, and gene expression analysis using overlapping feature selection methods. Sci Rep 11, 13323 (2021). <https://doi.org/10.1038/s41598-021-92725-8>
3. https://www.cancer.gov/pediatric-adult-rare-tumor/rare-tumors/rare-kidney-tumors/clear-cell-renal-cell-carcinoma#:~:text=Clear%20cell%20renal%20cell%20carcinoma%2C%20or%20ccRCC%2C%20is%20a%20type,called%20conventional%20renal%20cell%20carcinoma