Computational Analysis of Lung and Kidney Cancer gene expressions and their CNV relationship.

Mohamed El-Manzalawi, Amany Awad , Rania Alanany, Ahmed Elghamry

School of Information Technology and Computer Science, Nile University, Egypt

Course Code: CIT660: Statistical Analysis and Visualization

Under supervision of

Prof. Ibrahim Mohamed Youssef, PhD

Systems and Biomedical Engineering,

Faculty of Engineering, Cairo University

Abstract

Lung and kidney cancer are both one of the leading causes of cancer-related deaths around the world. For Kidney cancer it is the 13th most common cancer worldwide, accounting for 2.4% of all cancers, with more than 330,000 new cases diagnosed yearly[1]. As for lung cancer deaths are estimated at 135,000 patients per year which have become more numerous than the deaths from prostate, breast, brain, and colorectal cancer combined [2]. Therefore, in this study we aim to analyze the gene expression of different genes collected from cancer and healthy patients to determine most deferentially expressed genes, and identifying the CNVs that has the largest effect on those genes. The cancer types are: Lung Squamous Cell Carcinoma (LUSC) and Kidney Renal Clear Cell Carcinoma (KIRC). Moreover doing GSEA on the DEGs to identify upregulated and down regulated genes in both cancer cases. Using our R script, we will use the data we have to identify the most deferentially expressed genes based on the test statistic value. Then we will choose the top 5 genes and preform multi variable regression analysis on them to determine which CNV affect each gene. After that we will perform GSEA on the genes using the hallmarks gene set. This process will be repeated on both cancer types. After observation of the results various Genes showed significant gene expression such as TMEM133 in kidney and FGF11 in lung which might be used as way to detect cancer upon further future studies and up regulation in many genes related to cell division.

Keywords: Lung Cancer, Kidney Cancer, Gene Expression, GSEA, CNV.

Member Contributions	
Ahmed Elghamry	Equally contributed the codes of data filtration, reading data files, introduction, CNV filtration codes, literature review and presentation.
Rania Alanany	
Amany Awad	Equally contributed the rest of tasks (Codes and report for hypothesis testing and regression).
Mohamed El-Manzalawy	

1. Introduction

Lung cancer is the leading cause of cancer-related death worldwide. Despite the advances in treatment methods that have been made available in recent years, including minimally invasive surgical approaches, chemotherapies, and targeted therapies, the 5-year survival of patients with lung cancer is ranging between 10% and 20% for most geographic areas [3]. Both genetic and environmental factors contribute to the development of lung cancer. Eighty percent of lung cancers develop in current or former smokers with perhaps an additional 5% arising from passive tobacco smoke exposure. Additional risk factors include asbestos, radon, and radiation exposure. Susceptibility to lung cancer may be determined in part by the capacity to activate and detoxify inhaled procarcinogens from the environment. There are four major histological types of lung cancer: SCLC, adenocarcinoma, squamous cell carcinoma, and large cell carcinoma [4]. Squamous cell carcinoma (SCC) is a major histologic type and comprises ~30% of all pulmonary tumors. SCC is characterized by the presence of cytoplasmic keratinization and/or desmosomes (intracellular bridges). Clinically, SCC tumors occur more often in smokers and males compared with the other histologic types. Patients affected with SCC tumors show a wide range of clinical outcomes. For instance, 83% of autopsied SCC patients had regional metastases and 68% of SCC stage I patients survived beyond 5 years [5].

Kidney Renal Clear Cell Carcinoma (KIRC) is the eighth most common cancer and is known to be the most lethal of all the genitourinary tumors with an estimation of approximately 65,000 new cases and approximately 13,000 deaths annually in United States. This disease is known resistant to radiotherapy and chemotherapy, there are very few cases that have been reported to respond immunotherapy. If KIRC can be detected in very early stages, it is potentially curable by surgical resection, while adjuvant therapies have not been proven beneficial. The recurrence rate is not very high, although still considered not uncommon. Nevertheless, there is no curative treatment for late stage KIRC. The 2-year survival rate of patients with metastatic KIRC is less than 20%. Therefore, further investigations of the genomic alterations and underlying molecular mechanisms are essential for early diagnosis and treatment [6].

As cancer is a consequence of the accumulation of genetic alterations and dysregulation of pathways, identification of differentially expressed genes and pathways is important. We aimed to develop integrative approaches to identify differentially expressed genes. In this study, we

designed computational approaches to identify differentially expressed genes for data provided by the TCGA data portal (The Cancer Genome Atlas).

2. Methods

2.1. Hypothesis testing

The data for each cancer type are GE for tissues with cancer (KIRC and LUSC) and for tissues in a healthy case. For cases (patients), each GE is described from two tissues from the same individuals. Copy number alterations (CNAs) for cancer are described for chromosome segments: arm-level and focal-level.

To identify the DEGs for each cancer type, we used the hypothesis testing, Fold Change, and volcano plot methods. For the hypothesis testing method, we considered the samples once are paired and once are independent. For the volcano plot method, we used the set of DEGs obtained by the hypothesis that samples are paired only. In the two cases of paired and independent samples, we checked the normality of genes using Shapiro-test which showed that Wilcoxon-test should be used in the two cases.

GE Data Filtration

We would like to point out that the original gene expression data for all groups contained more than 50% zeros for many genes, which led to no results for these genes when we applied the Wilcoxon-test to the considered groups. This happened at the beginning of our study, so we filtered the original data by deleting all genes containing more than 50% zeros for each cancer type group and from the corresponding healthy group before carrying out our study. Table (1) shows the numbers of genes and samples for the original data and filtered data. For filtered groups, we used Wilcoxon signed rank test with the paired samples, while Wilcoxon rank sum test was used with the independent samples.

	Original data		Filtered data	
	Kidney	Lung	Kidney	Lung
No. of genes	19216	19648	17034	17284
No. of samples	68	50	68	50

Table 1. Numbers of genes and samples for the original data and filtered data.

2.2. GSEA

After extracting the DEGs for both cancer conditions based on test-statistic values we had to change the format to make it suitable for the GSEA tool. We constructed 2 files for each cancer type that had all the healthy and cancer genes expression together. Since we have a 68 samples in both cases of lung cancer the files had total of 136 samples. Moreover, in order to prevent errors in the GSEA tools caused by the samples having the same name, the names of cancer samples were changed by adding "C-"at the start of each sample using a R Code shown in the figure (1).The files were saved in a text tab delimited format with **".gct"** extension as it is required for the GSEA tool to run.

```
path="File Path"
GE.Data <- read.table(path, sep = ',')
View(GE.Data)
for (i in 2:ncol(GE.Data)) {
  Original=GE.Data[1,i]
  Mod=paste("C-",Original)
  GE.Data[1,i]=Mod
}
path.name.out_2 <- gsub("Lung_Degs_Paired_Cancer.csv","Filtered_Lung_Cancer.csv",path)
write.table(GE.Data, path.name.out_2,sep=",")
```

Figure 1. Shows The R Script used in changing the names of cancer samples [Lungs].

For the GSEA to run we needed to create another file for both cancer types.it only required to define the groups we are using, the number of genes present in the analysis and the number of samples. It had to be saved in a tab delimited format with **".cls"** extension. Then we loaded the data to the GSEA tool setting our gene set form the gene set database to hallmarks gene set.

CNV Data Filtration

After observing the CNV data we noticed that there were numerous values that had zeros. So we started doing some filtration taking into consideration the same threshold used in our previous data filtration as to delete CNVs that has more than 50% zeros in its values. Some columns had "NA" values so we decided to remove them as well. This could be maybe due to error in the methods used to calculate the CNVs or a random one during data extraction. Either way it had to be removed in order to get more accurate results. This was done for both cancer types. The R script we used is shown in figure (2) below.

```
cnv_delete3=c()
R_K=nrow(kirc_cnv)/2
for (i in 1:ncol(kirc_cnv)){
  if (length(which(kirc_cnv[,i]==0)) >= R_K ) {
    cnv_delete3=append(cnv_delete3,i)
  }
}
kirc_cnv=kirc_cnv[,-cnv_delete3]
K_Column_NA=names(which(colSums(is.na(kirc_cnv))>0))
Delete_K=which(colnames(kirc_cnv)==K_Column_NA)
kirc_cnv=kirc_cnv[,-Delete_K]
```

Figure 2. Shows The R Script used in filtering the CNV data of cancer samples [Kidney].

2.3. Regression

We collected the most significant expression genes from the results of our hypothesis testing and created a data frame. Since the samples in the CNV are different from the samples in the gene expression data we used the function intersect and extracted the common samples between both files. But since the names are also different as there was a "." In the sample names in CNV file and "-"in the gene expression files we had to convert one of them to the other so we changed any "." In the names of the samples and started the intersection function. The code is shown in figure (3).

```
five_most_expressed_genes_Kidney <- kirc_t[intersect(row.names(Top_5_Kidney),rownames(kirc_t)),]

for (i in 1:nrow(kirc_cnv)){
  kirc_cnv[i,1]=gsub("-",".",kirc_cnv[i,1])
}
rownames(kirc_cnv)=kirc_cnv[,1]
kirc_cnv=kirc_cnv[,-1]

cnv_of_five_most_expressed_Kidney = kirc_cnv[intersect(colnames(five_most_expressed_genes_Kidney),rownames(kirc_cnv)),]
cnv_of_five_most_expressed_Kidney <- as.matrix(cbind(cnv_of_five_most_expressed_Kidney))
```

Figure 3. Shows the code used in getting the common samples after modifying the names.

Then we started our regression on each gene and printing the results in separate files making it easier to interpret each gene result. However the regression in lung presumed little bit difficult as the predictors (CNVs) were bigger then the data points so we had to penalize the CNVS based on their estimated effect. This was done using the "glmnet" package. After identifying the CNVs we ran each gene's regression with their corresponding CNVs. The code for this process is shown below in figure (4).

```
library(glmnet)
CNV_Accepted <- list()
Variables_Number <- dim(cnv_of_five_most_expressed_Lung)[2]
for ( i in 1:nrow(five_most_expressed_genes_Lung)){
  fit_cv <- cv.glmnet(cnv_of_five_most_expressed_Lung, five_most_expressed_genes_Lung[i,], family="gaussian", alpha=1, standardize=FALSE, nfolds=5)
  lambda <-fit_cv$lambda.min
  model <- glmnet(cnv_of_five_most_expressed_Lung, five_most_expressed_genes_Lung[i,],alpha=1, lambda=lambda, standardize=FALSE)
  coef_fit <- coef(model, s=lambda)[2:(Variables_Number+1)]
  CNV_Accepted[[i]] <- which(abs(coef_fit) > 0)
}
```

Figure 4. Shows the use of glmnet package to penalize the CNVs.

2.4. Enrichment of functional terms and KEGG pathways within kidney and lung cancer DEGs:

We performed the kidney cancer and lung cancer DEGs into DAVID network (https://david.ncifcrf.gov/summary.jsp) to get the Gene Ontology (GO Term) (Biological process (BP), Cellular compartment (CC), and Molecular Function (MF)), OMIM disease and KEGG pathways.

3. Results and Discussion

3.1. Hypothesis testing

For the hypothesis testing method, we identified the set of DEGs then we corrected it using the False Discovery Rate (FDR) method (Benjamini-Hochberg procedure). The set of DEGs for the two types (KIRC and LUSC) are reported in the following:

- DEGs by the hypothesis testing method are 13008 genes for KIRC paired groups.
- DEGs by the Fold Change method are 7286 genes for KIRC paired groups.
- DEGs by the hypothesis testing method are 13148 genes for LUSC paired groups.
- DEGs by the Fold Change method are 9141 genes for LUSC paired groups.
- DEGs by the hypothesis testing method are 12848 genes for KIRC independent groups.
- DEGs by the hypothesis testing method are 13240 genes for LUSC independent groups.

The set of DEGs show that the expression level differs under healthy and cancerous tissues for most genes in the two types (KIRC and LUSC) whether the samples are paired or independent. The volcano plots in figures (5, 6) show the Log2FoldChange, corrected p-values, and DEGs for KIRC and LUSC groups.

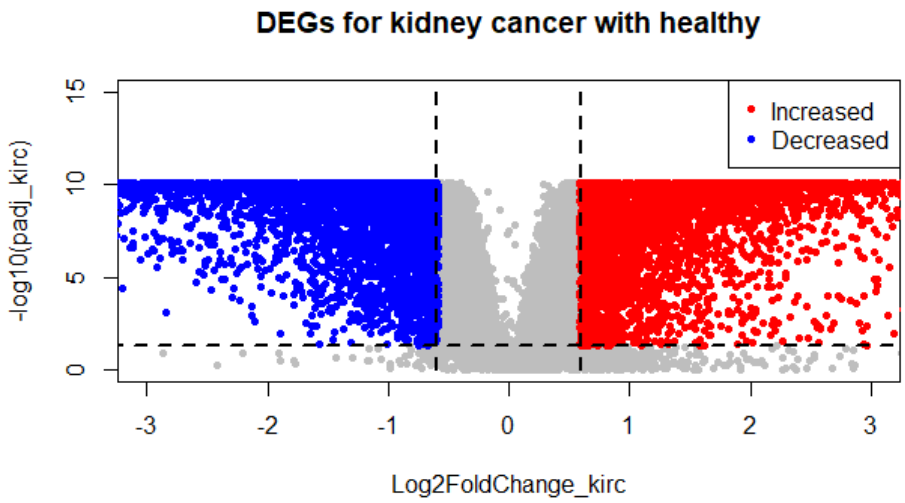

Figure 5. Volcano plot shows the Log2FoldChange, corrected p-values, and DEGs for KIRC *groups.*

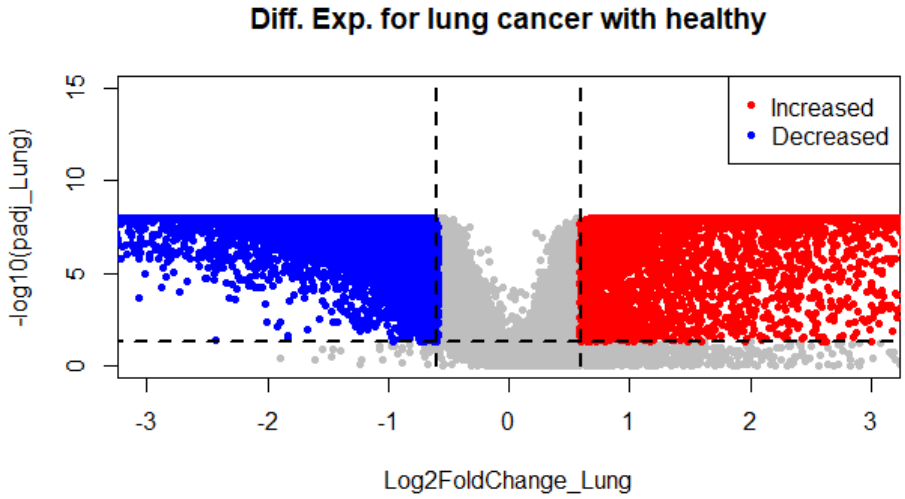

Figure 6. Volcano plot shows the Log2FoldChange, corrected p-values, and DEGs for LUSC *groups.*

For the volcano plots, the set of DEGs are plotted by both of the hypothesis testing and Fold Change methods together. In Figures 5 and 6, the set of up-regulated genes are described by the red points while the down-regulated genes are described by the blue points. They are reported in the following:

- The up-regulated genes are 3550 genes for KIRC paired groups.
- The down-regulated genes are 3417 genes for KIRC paired groups.
- The up-regulated genes are 4037 genes for LUSC paired groups.
- The down-regulated genes are 4522 genes for LUSC paired groups.

According to Venn diagram we make two pairwise comparisons, including different DEGs for kidney in two cases (dependent and paired). The total number of significant kidney genes (n=13310) with 462 genes unique for paired case, 302 genes for independent case and 12546 genes are common for both cases (fig7). On the other hand the total number of significant lung genes (n=13715) with 475 genes unique for paired case, 567 genes for independent case and 12673 genes are common for both cases (fig8).

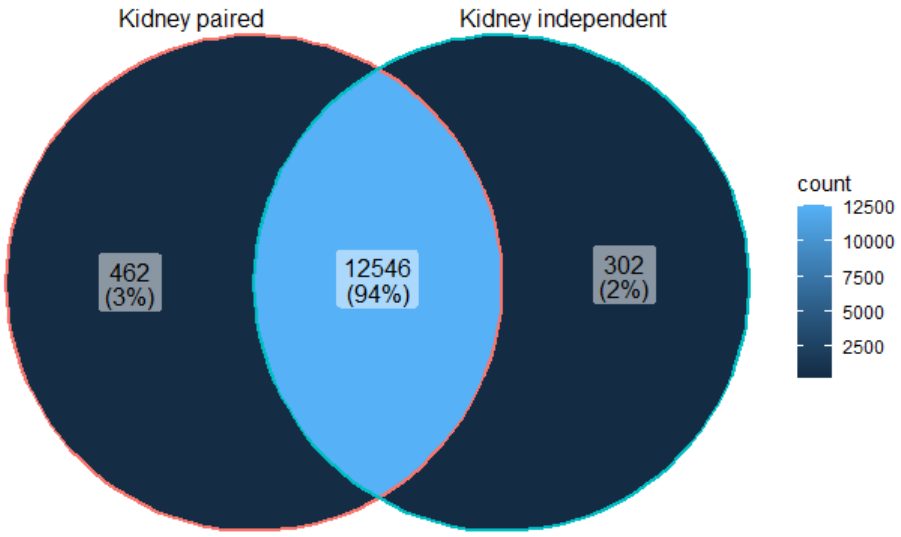

Figure 7. Venn diagram for DEGs of kidney in the two cases (paired and independent).

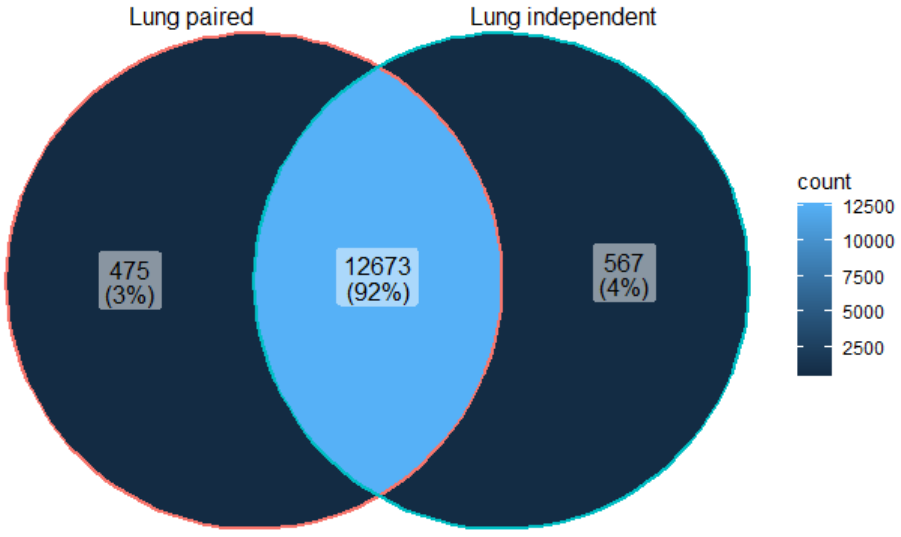

Figure 8. Venn diagram for DEGs of lung in the two cases (paired and independent).

3.2. GSEA

Lung Cancer

The GSEA showed the in lung cancer 16 / 50 gene sets are upregulated in Cancer and 34 / 50 gene sets are upregulated in Control. The most significant gene sets in Cancer was Gene Set:

(1) HALLMARK_E2F_TARGETS which are E2Fs are critical regulators of the cell cycle.

(2) HALLMARK_G2M_CHECKPOINT and

(3) MYC_TARGETS_V1 also used in progression through the cell division cycle. All 3 were mostly affected by KIF4A and KPNA2 gene .The GSEA for all 3 gene sets are shown below (figures 9, 10, 11).

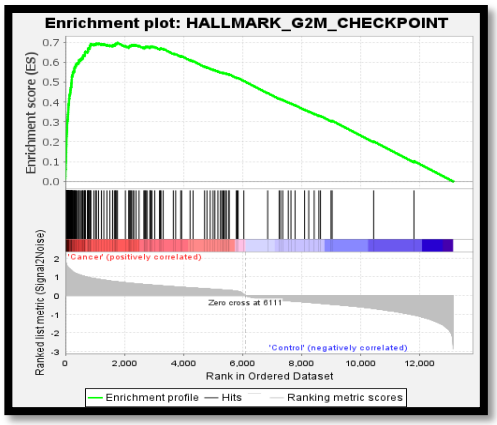

Figure 9. G2M upregulation in

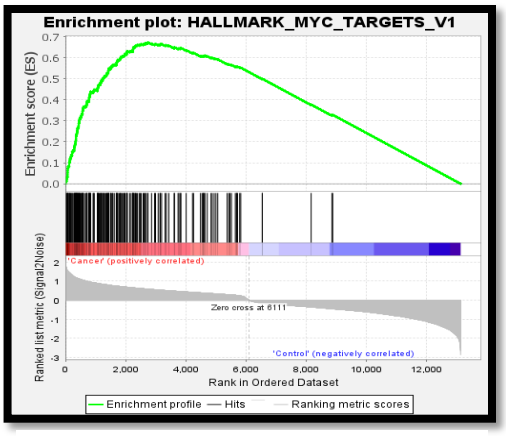

Figure 10. MYC upregulation in

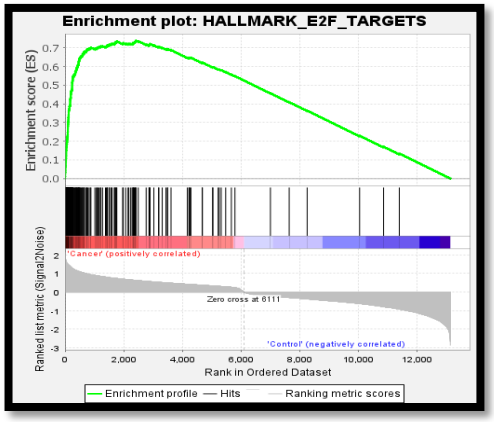

Figure11. E2F upregulation in

Kidney Cancer

The GSEA results showed 32 / 50 gene sets are upregulated in Cancer and 18 / 50 gene sets are upregulated in phenotype Control. The most significant gene sets in Cancer was Gene Set:

(1) HALLMARK_ALLOGRAFT_REJECTION,

(2) HALLMARK_INTERFERON_GAMMA_RESPONSE

which correlates with studies that presume that interferon gamma has role in promoting cancer[3].

(3) HALLMARK_INTERFERON_ALPHA_RESPONSE.

The GSEA for all 3 gene sets are shown below (figures 12, 13, 14).

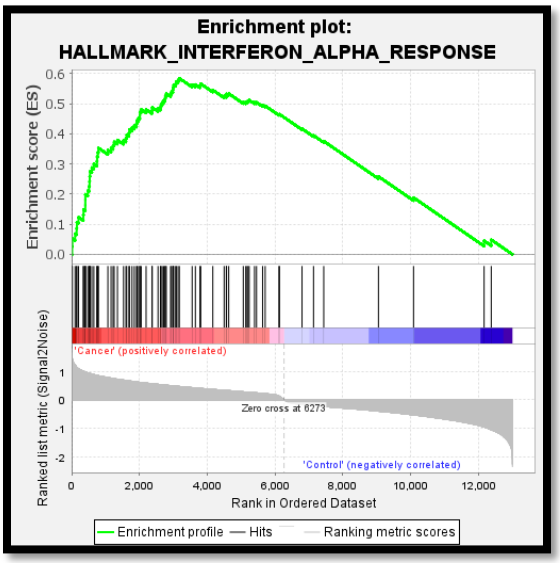

Figure 12. INF Alpha upregulation in cancer.

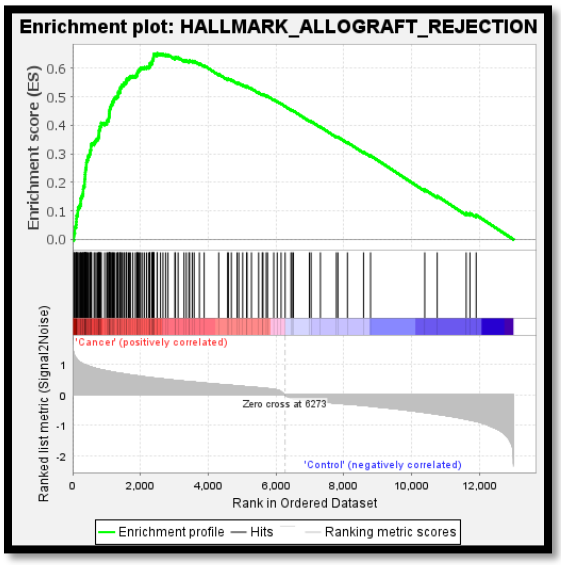

Figure 13. Allograft rejection upregulation in cancer.

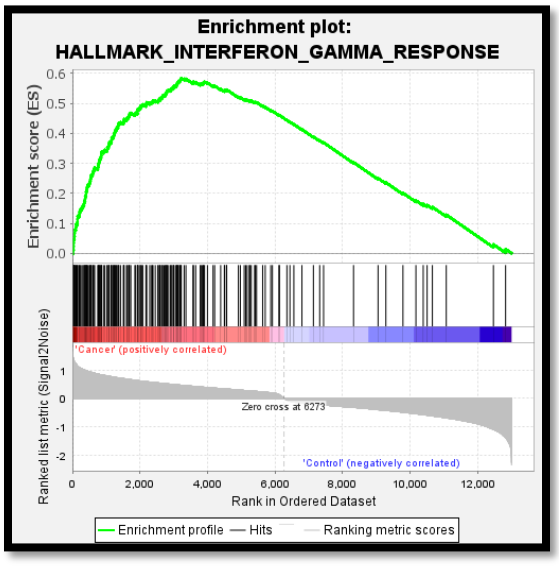

Figure 14. INF γ upregulation in cancer.

3.3. Regression

Lung Cancer

Only 3 from the most significant genes in lung cancer showed that it can be affected by CNVs.The genes are (1) FBXO45 which highly affected by CNV_11q13.3, (2) TOMM70A which highly affected by yCNV_8p11.23, and (3) WDR53 which was affected by yCNV_3q26.33 .The Figure (15) below shows one of the result of the lung cancer.

```
[1] "The CNV are :"
             Estimate Std..Error   t.value     Pr...t..
(Intercept)  724.4007   60.26793 12.019671 2.876751e-07
yCNV_8p11.23 173.5126   52.19817  3.324112 7.693892e-03

Call:
lm(formula = x ~ y)

Residuals:
    Min      1Q  Median      3Q     Max
-212.71 -114.17  -12.99  100.93  289.95

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    724.40      60.27  12.020 2.88e-07 ***
yCNV_11q13.3   -45.77      37.59  -1.217  0.25138
yCNV_8p11.23   173.51      52.20   3.324  0.00769 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 172 on 10 degrees of freedom
Multiple R-squared:  0.6144,    Adjusted R-squared:  0.5372
F-statistic: 7.965 on 2 and 10 DF,  p-value: 0.008529
```

Figure 15. Shows regression results of TOMM70A gene.

Kidney Cancer

As for kidney cancer also 3 genes were found to relate with the CNV (1) SND1 affected by yCNV_9p21.3, yCNV_5q35.1, and yCNV_9p23. (2) DAGLB and it is corresponding CNV yCNV_7q36.3. (3) SERPINH1 affected by several CNVs yCNV_8p23.2, yCNV_9p21.3, and yCNV_Xq11.2.Part of the results of one of the kidney genes is shown below.

```
[1] "The significant CNV are :"
             Estimate Std..Error   t.value     Pr...t..
(Intercept)  7769.925   1308.681  5.937217 2.167112e-06
yCNV_Xq11.2 14904.429   7212.075  2.066594 4.813277e-02
yCNV_8p23.2 -4366.134   1748.280 -2.497388 1.866249e-02
yCNV_9p21.3 -9824.977   3871.231 -2.537946 1.699830e-02
[1] "The Full results of the regression:"

Call:
lm(formula = x ~ y)

Residuals:
    Min      1Q  Median      3Q     Max
-2926.7 -1121.5   -89.1  1046.1  4750.7

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     7769.9     1308.7   5.937 2.17e-06 ***
yCNV_Xq11.2    14904.4     7212.1   2.067   0.0481 *
yCNV_17q24.3    1254.8     5849.3   0.215   0.8317
yCNV_7q36.3     1142.9     1790.3   0.638   0.5284
yCNV_3p21.32   -4953.4     4435.0  -1.117   0.2735
yCNV_14q31.1     928.6     1797.8   0.517   0.6095
yCNV_3q26.32    -275.1     2135.8  -0.129   0.8984
yCNV_8p23.2    -4366.1     1748.3  -2.497   0.0187 *
```

Figure 16. Shows regression results of SERPINH1 gene.

3.4. David enrichment analysis

We found that apoptotic process and signal transduction are the most significant biological process (BP) which are highly enriched with our list of kidney DEGs while protein phosphorylation and positive regulation of transcription from RNA polymerase II premotor in list of lung DEGs. The following cellular components (CC Go-terms) are the most significant term which highly enriched in our list of kidney and lung DEGs is cytosol also, in (MF Go-terms) the most significant term which highly enriched in our list of kidney and lung DEGs is protein binding in the two cases. More over there is a close relation between Leukemia, acute myeloid, somatic and breast cancer in our interested list of DEGs in lung cancer. The KEGG map pathway in cancer indicates genes that our two lists of DEGs have been enriched in kidney and lung cancer (supplementary files).

4. Literature review

By literature review we found that some genes from the top five differentially expressed genes are associated with lung and kidney cancer.

(a) FBXO45 is a novel biomarker for the diagnosis and treatment of patients with SqCLC, it is up-regulated and involved in the tumorigenesis of SqCLC [8].

(b) FGF11(fibroblast growth factor 11) is up-regulated in NSCLC. It's expression is associated with poor prognosis. FGF11 knockdown inhibited proliferation, invasion and migration of NSCLC and supressed tumor growth, while its overexpression is promoted tumor growth [9].

(c) SERPINH1 is prognostic marker in ccRCC, it is is up-regulated in a grade-dependent manner. So, the high level of SERPINH1 has strong association with poor prognosis of ccRCC patients [10].

(d) SND1 mRNA expression is significantly up-regulated in ccRCC tissues [11].

References

[1] International Agency for Research on Cancer: GLOBOCAN 2012: Estimated cancer incidence, mortality and prevalence worldwide in 2012
v1.0. http://publications.iarc.fr/Databases/Iarc-Cancerbases/GLOBOCAN-2012-Estimated-Cancer-Incidence-Mortality-And-Prevalence-Worldwide-In-2012-V1.0-2012

[2] Siegel R. L, Miller K. D., Jemal A. Cancer statistics, 2020. CA Cancer J Clin. (2020); 70(1):7-30.

[3] Yang Y, Liu B. Exploring and comparing of the gene expression and methylation differences between lung adenocarcinoma and squamous cell carcinoma, J Cell Physiol, (2019); 234(4):4454-4459. doi: 10.1002/jcp.27240.

[4] Amy L. McDoniels-Silvers; Caramella F. Nimri; Gary D. Stoner; Ronald A. Lubet; Ming You, Differential Gene Expression in Human Lung Adenocarcinomas and Squamous Cell Carcinomas, Clin Cancer Res (2002) 8 (4): 1127–1138.

[5] Wilkerson, M. D.; Yin, X.; Hoadley, K. A.; Liu, Y.; Hayward, M. C.; Cabanski, C. R.; Muldrew, K.; Miller, C. R.; Randell, S. H.; Socinski, M. A.; Parsons, A. M.; Funkhouser, W. K.; Lee, C. B.; Roberts, P. J.; Thorne, L.; Bernard, P. S.; Perou, C. M.; Hayes, D. N. Lung Squamous Cell Carcinoma mRNA Expression Subtypes Are Reproducible, Clinically Important, and Correspond to Normal Cell Types. Clinical Cancer Research, (2010) 16(19), 4864–4875. doi:10.1158/1078-0432. CCR-10-0199

[6] Yang W., Yoshigoe K., Qin X., Liu J. S., Yang J. Y., Niemierko A., Deng Y., Liu Y., Dunker A. K., Chen Z., Wang L., Xu D., Arabnia H. R., Tong W. & Yang M. Q., Identification of genes and pathways involved in kidney renal clear cell carcinoma, BMC Bioinformatics, 15: S2 (2014).

[7] Mojic M, Takeda K, Hayakawa Y. The Dark Side of IFN-γ: Its Role in Promoting Cancer Immunoevasion. Int J Mol Sci. (2017);19(1):89. doi: 10.3390/ijms19010089.

[8] Wang K., Qu X., Liu S.,Yang X., Bie F., Wang Y., Huang C., Du J. Identification of aberrantly expressed F-box proteins in squamous-cell lung carcinoma, Journal of Cancer Research and Clinical Oncology (2018) 144:1509–1521

[9] Xiaowei wu et al. Fibroblat growth factor 11 (FGF11) promotes non small cell lung cancer (NSCLC) Progression by regulating hypoxia signaling pathway, Journal of translational medicine (2021) 19(1), 1-14.

[10] Qi Y., Zhang Y., Peng Z., Wang L., Wang L., Feng D., He J., Zheng J., SERPINH1 overexpression in clear cell renal cell carcinoma: association with poor clinical outcome and its potential as a novel prognostic marker, J. Cell. Mol. Med. (2018) 22(2), 1224-1235 .

[11] He A., He S., Huang C., Chen Z., Wu Y., Gong Y., Li X., Zhou L., MTDH promotes metastasis of clear cell renal cell carcinoma by activating SND1-mediated ERK signaling and epithelial-mesenchymal transition, AGING (2020), 12(2) :1465-1487. doi: 10.18632/aging.102694.