

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

By

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

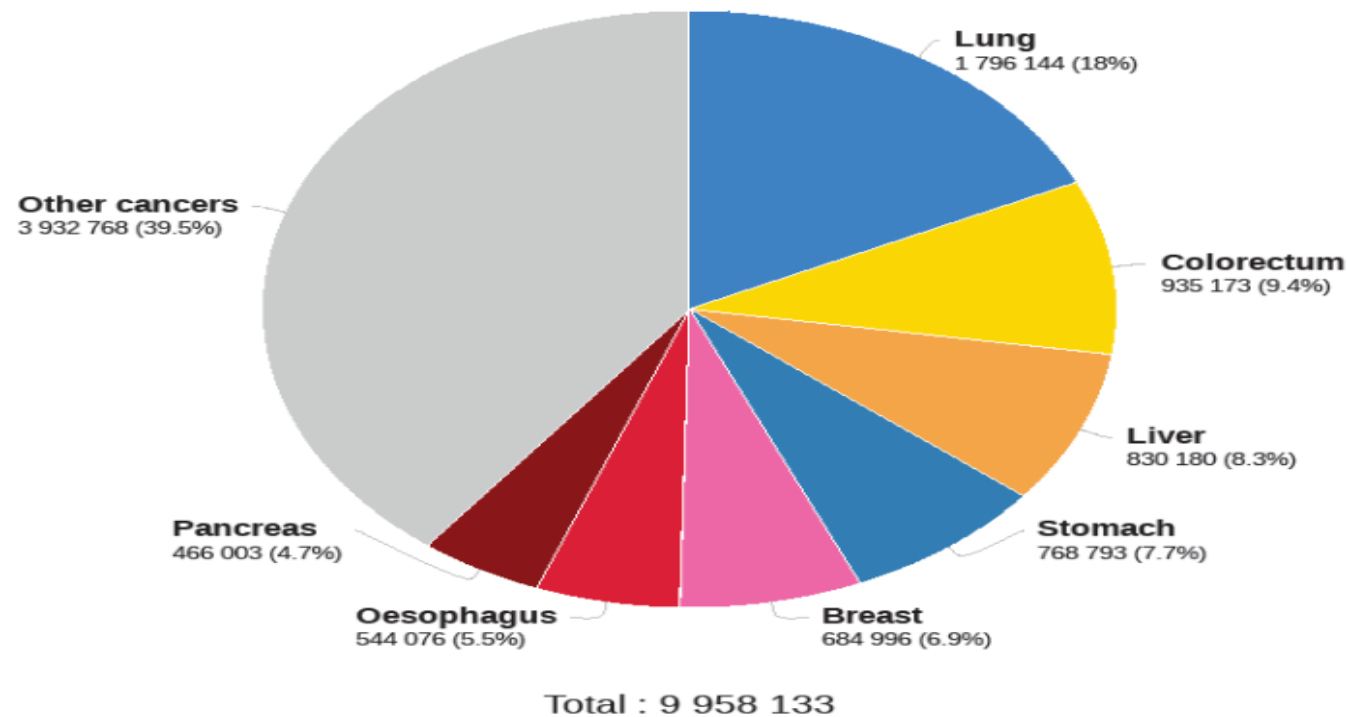
Contents

- **Introduction**
- **Methods**
- **Results**
- **Conclusion**

Introduction



Estimated number of deaths in 2020, worldwide, both sexes, all ages



Kidney and renal pelvis

AT A GLANCE

Estimated new cases,
2022

79,000

Estimated deaths, 2022

13,920

Incidence rates, 2014-
2018

17.1

Average annual rate per 100,000,
age adjusted to the 2000 US
standard population.

Death rates, 2015-2019

3.6

Average annual rate per 100,000,
age adjusted to the 2000 US
standard population



TCGA

The Cancer Genome Atlas Program



TCGAbiolinks

platforms all rank 87 / 2140 support 5 / 1 5 in Bioc 7 years
build error updated < 3 months dependencies 113

DOI: [10.18129/B9.bioc.TCGAbiolinks](https://doi.org/10.18129/B9.bioc.TCGAbiolinks)  

TCGAbiolinks: An R/Bioconductor package for integrative analysis with GDC data

Codes

Methods and Results

1- Data Filtration

- The original gene expression data for all groups **contained more than 50% zeros** for many genes, **led to no results** for these genes when we applied the Wilcoxon-test
- 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

1- Data Filtration

Numbers of genes and samples for the original data and filtered data

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

- Wilcoxon signed rank test with the paired samples
- Wilcoxon rank sum test was used with the independent samples.

1- Data Filtration

Filtration codes

```
rows_to_delete=c()
for (i in 1:nrow(kirc_t)){
  if (length(which(kirc_t[i,]==0)) >=ncol(kirc_t)/2 ) {
    rows_to_delete=append(rows_to_delete,i)
  } else if (length(which(kirc[i,]==0)) >= ncol(kirc)/2 ){
    rows_to_delete=append(rows_to_delete,i)
  }
}
kirc_t=kirc_t[-rows_to_delete,]
kirc=kirc[-rows_to_delete,]
```

This code is applied to the other groups.

Hypothesis testing

First case: Paired samples

Step 1

- Does the paired difference follow the normal distribution?

Step2

- We calculated the differences between the GE of Kidney cancer group and the corresponding healthy group.

Step3

- We wrote a function in a separated R.file.
- Its input = any data frame **(df)**,
- Output = vector of the p-values of Shapiro-test for all genes.

Paired samples

Step 4

- We called this function in the main R.script and applied it 6 times
- 2-times on difference between GE in the paired case (1-Kidney & 1-Lung).
- 4-times on each group (1 by 1) in the independent case (2-Kidney & 2-Lung).

```
shapiro.pvalues = function(df){Shapiro_P=c()  
n1= nrow(df)  
for(i in 1:n1){Shapiro_P[i] = shapiro.test(as.numeric(df[i, ]))$p.value}  
return(Shapiro_P)}
```

Paired samples

Step 5

We wrote code in R to check if there is Shapiro p-values < 0.05 (i.e. doesn't follow ND) we apply **Wilcoxon-test**. and if there is no, we apply **t-test**.

The output of this code is **Wilcoxon test should be applied**.

Step 6

- We applied **Wilcoxon signed rank test** (paired), and calculated two vectors:
- Vector1= p-values test, Vector2= the statistic test (calculate top 5 sig. genes)

```
Wilcox_P = c()
Wilcox_T = c()
n1 = nrow(kirc_t)
for (i in 1:n1){Wilcox_Kidney = wilcox.test(x = as.numeric(kirc_t[i, ]), y = as.numeric(kirc[i, ]),
                                           alternative = 'two.sided', paired = TRUE )

Wilcox_P[i] = Wilcox_Kidney$p.value
Wilcox_T[i] = Wilcox_Kidney$statistic }
```

Paired samples

Step 7

- We calculated the adjusted P-values

```
padj_kirc=p.adjust(Wilcox_P, method = 'fdr')
```

Step 8

- We calculated the Log2Fold Change between Kirc_t & Kirc groups.

```
Log2FoldChange_kirc=log2(rowMeans(kirc_t)) - log2(rowMeans(kirc))
```

Step 8





- We calculated the **DEGS**, which are genes of $|\text{Log2FoldChange}| > \log_2(1.5)$

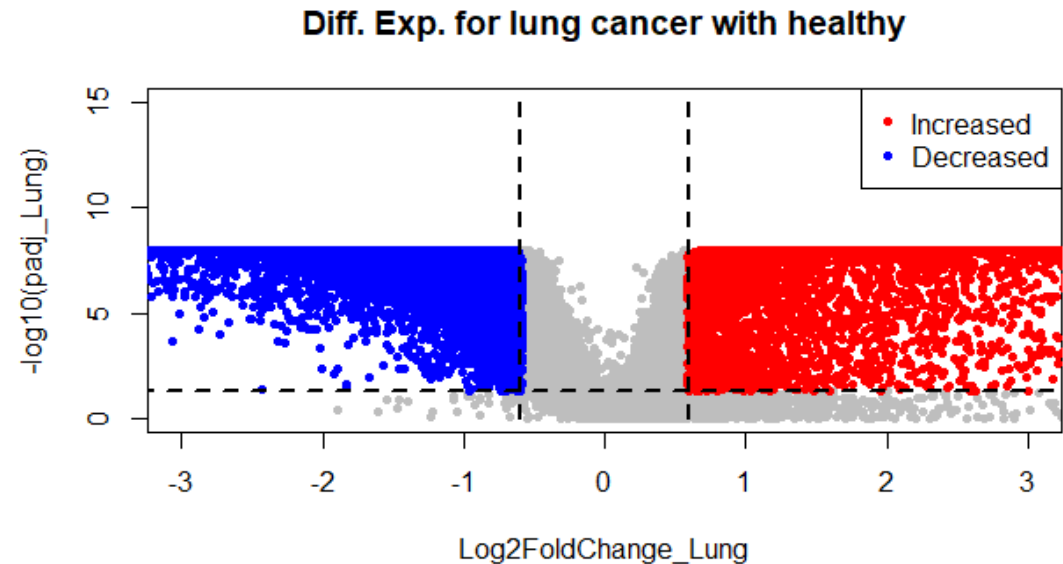
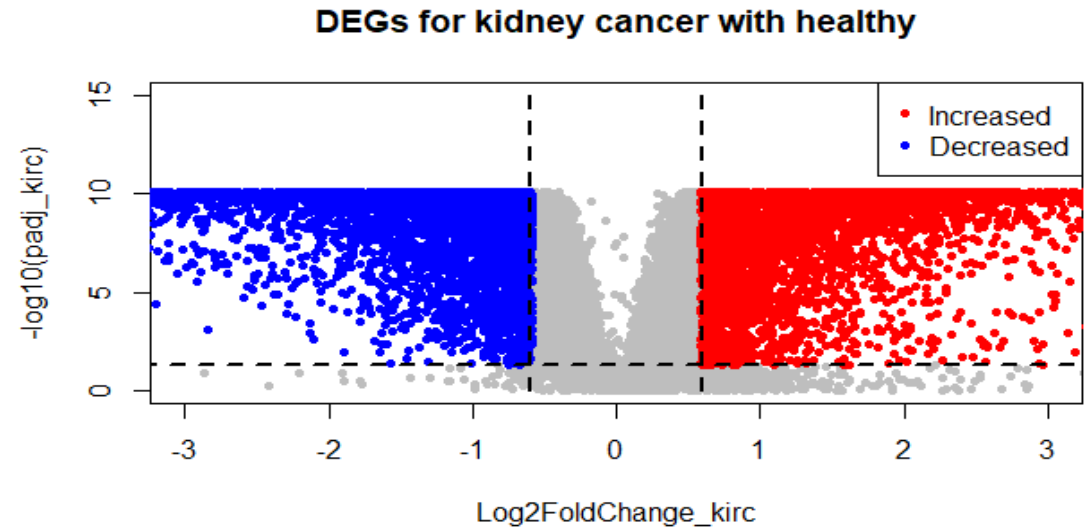
DEGs	KIRC paired groups	LUSC paired groups
Hypothesis testing	13008 genes	13148 genes
Fold Change method	7286 genes	9141 genes

The DEGs by the two methods (hypothesis , Log Fold)

Paired samples Volcano plot method

Volcano plot with the two conditions:

- **Up-regulated genes:** $\text{padj} < 0.05$ & $\text{Log2FoldChange} > \log_2(1.5)$
- **Down-regulated genes:** $\text{padj} < 0.05$ & $\text{Log2FoldChange} < -\log_2(1.5)$
- KIRC paired groups:
3550 genes  and 3417 genes 
- LUSC paired groups:
4037 genes  4522 genes 



Paired samples

The 5 most significant genes

Kidney

"SERPINH1" "DAGLB" "TMEM133" "FRMD8"
"SND1"

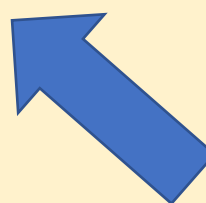
Lung

"WDR53" "FGF11" "TOMM70A" "FBXO45"
"GTF2IRD1"

Second case: Independent samples

We followed the same above steps but with **Wilcoxon rank sum test** for independent samples

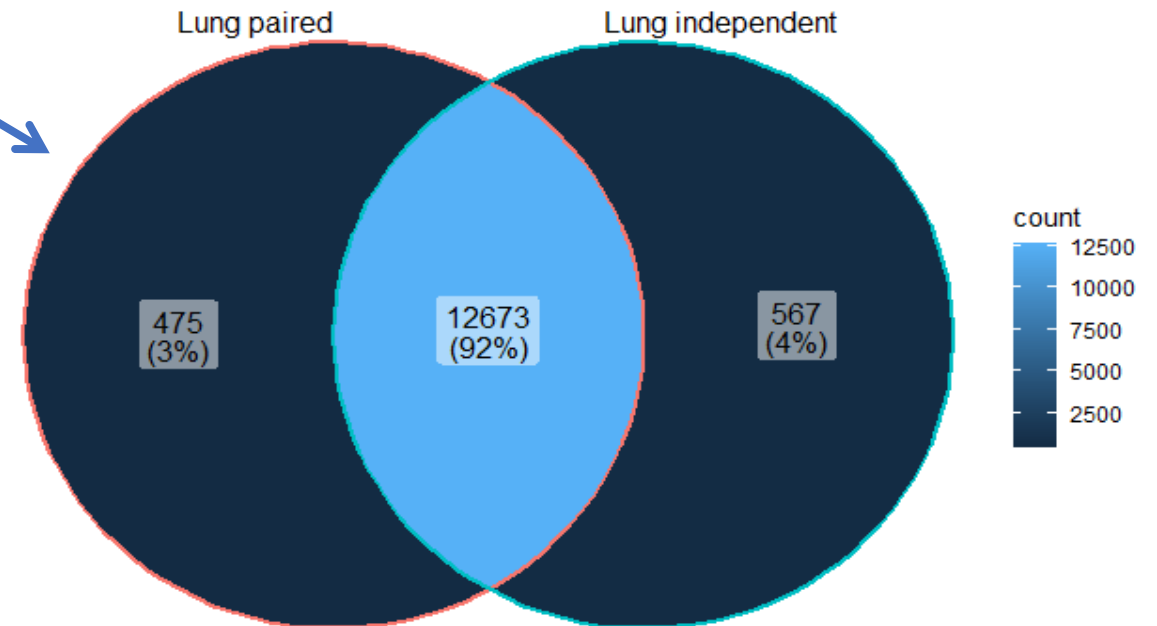
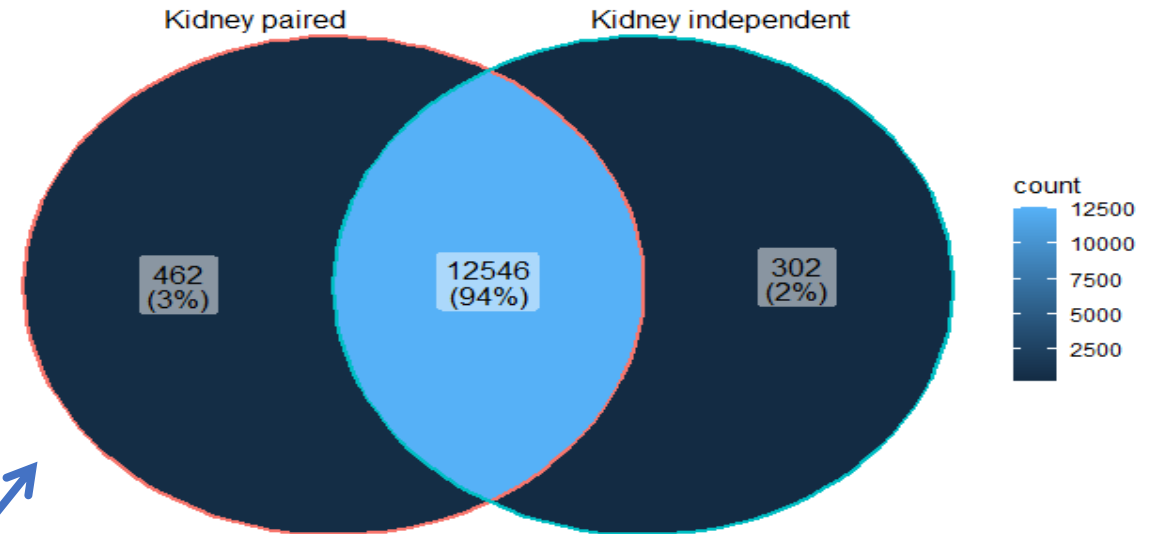
```
Wilcox_Inde_P = c()  
Wilcox_Inde_T = c()  
n1 = nrow(kirc_t)  
for (i in 1:n1){Wilcox_Inde_Kidney = wilcox.test(x = as.numeric(kirc_Inde_t[i, ]),  
y = as.numeric(kirc_Inde[i, ]), alternative = 'two.sided', paired = FALSE)  
Wilcox_Inde_P[i] = Wilcox_Inde_Kidney$p.value  
Wilcox_Inde_T[i] = Wilcox_Inde_Kidney$statistic }
```



- 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

Independent samples

The set of DEGs show that the expression level differs under healthy and cancerous tissues for most genes in the two types (Kidney and Lung) whether the samples are paired or independent.



Regression & GSEA

Regression

Step 1

- Unify sample names in the CNV file gene expression data file by replacing any “.” In the names of the samples with “-”

Step2

- Use the function intersect and extracted the common samples between both files.

```
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))
```

Regression

Printing the results in separate files making it easier to interpret each gene result

```
# For loop on all genes and exporting it into separate files.

Gene_Names_K=rownames(five_most_expressed_genes_kidney)
for (i in 1:5) {
  sink(paste(path.name,"/", "kidney_", Gene_Names_K[i], "_regression.txt", sep = ""))
  Genes_K_Regression=linear_model(five_most_expressed_genes_kidney[i,], cnv_of_five_most_expressed_kidney)
  Genes_Coeff_K= data.frame(summary(Genes_K_Regression)$coefficients)
  print("The significant CNV are :")
  print(Genes_Coeff_K[Genes_Coeff_K[,4]<=0.05,])
  print("The Full results of the regression:")
  print(summary(Genes_K_Regression))
  sink()
  closeAllConnections()
}
print("Kidney Regression Done :D")
```

Regression

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

```
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)
}
```

- 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 GSEA results showed 32 / 50 gene sets are upregulated in Kidney Cancer and 18 / 50 gene sets are upregulated in phenotype Control

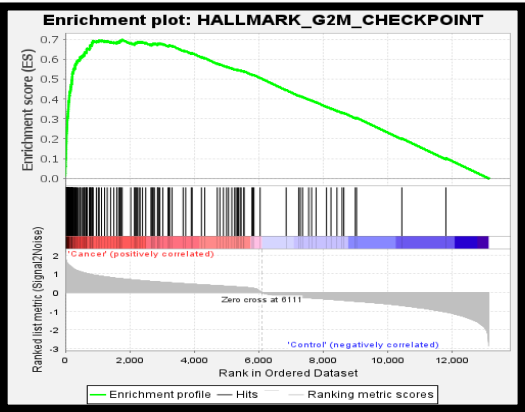


Figure 9. G2M upregulation in cancer.

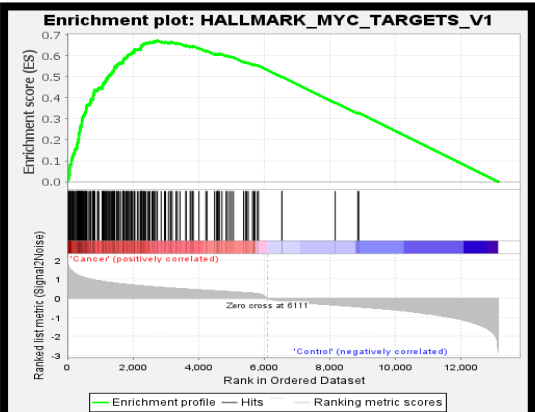


Figure 10. MYC upregulation in cancer.

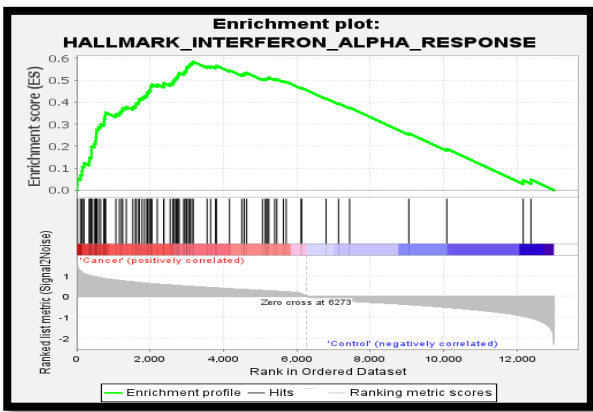


Figure 12. INF Alpha upregulation in cancer.

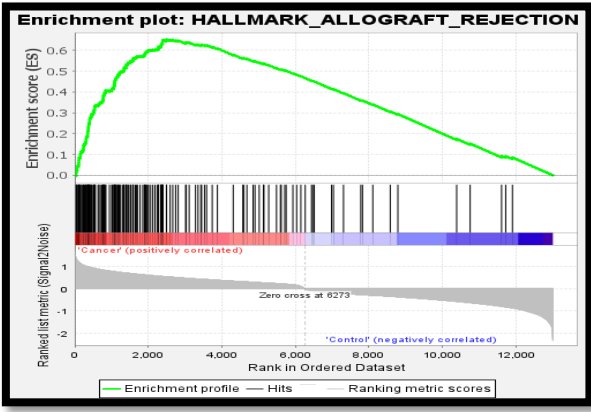


Figure 13. Allograft rejection upregulation in cancer.

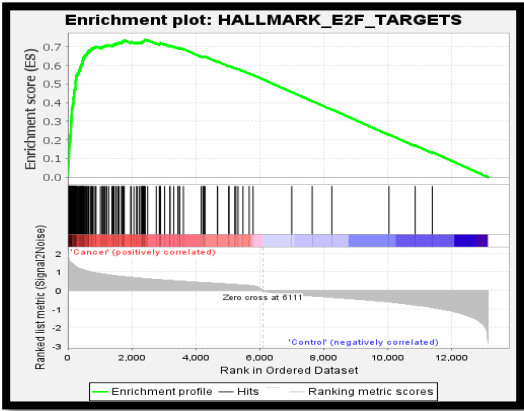


Figure11. E2F upregulation in cancer.

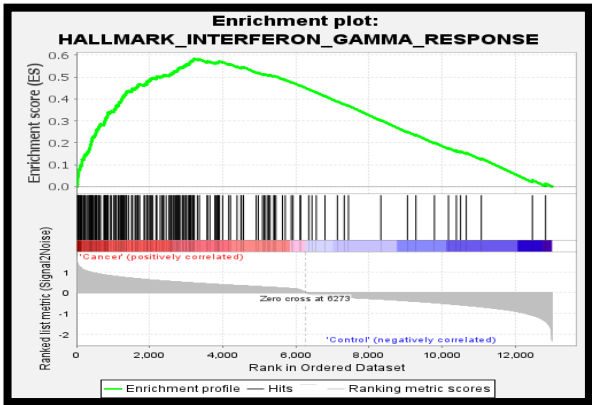


Figure 14. INF γ upregulation in cancer.

Regression & GSEA

- In 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

Regression results of SERPINH1 gene

```
[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 *

Regression & GSEA

- In lung cancer Only 3 genes from the most significant genes are affected by CNVs:

1- FBXO45 which highly affected by CNV_11q13.3

2-TOMM70A which highly affected by yCNV_8p11.23

3-WDR53 which was affected by yCNV_3q26.33

Regression results of TOMM70A gene

```
[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
```

Conclusion & Literature Review

CCRC

- **SND1 mRNA expression is significantly upregulated in ccRCC tissues**
- **SERPINH1 is prognostic marker in ccRCC, high level of SERPINH1 has strong association with poor prognosis of ccRCC patients**

MTDH promotes metastasis of clear cell renal cell carcinoma by activating SND1-mediated ERK signaling and epithelial-mesenchymal transition

SERPINH1 overexpression in clear cell renal cell carcinoma: association with poor clinical outcome and its potential as a novel prognostic marker

Lung Cancer

- **FBXO45 is a novel biomarker and upregulated in SqCLC.**
- **FGF11(fibroblast growth factor11) is upregulated in NSCLC. It's expression is associated with poor prognosis.**

Journal of Cancer Research and Clinical Oncology (2018) 144:1509–1521
<https://doi.org/10.1007/s00432-018-2653-1>

ORIGINAL ARTICLE - CLINICAL ONCOLOGY

Identification of aberrantly expressed F-box proteins in squamous-cell lung carcinoma

Wu et al. *J Transl Med* (2021) 19:353
<https://doi.org/10.1186/s12967-021-03018-7>

Journal of
Translational Medicine

RESEARCH

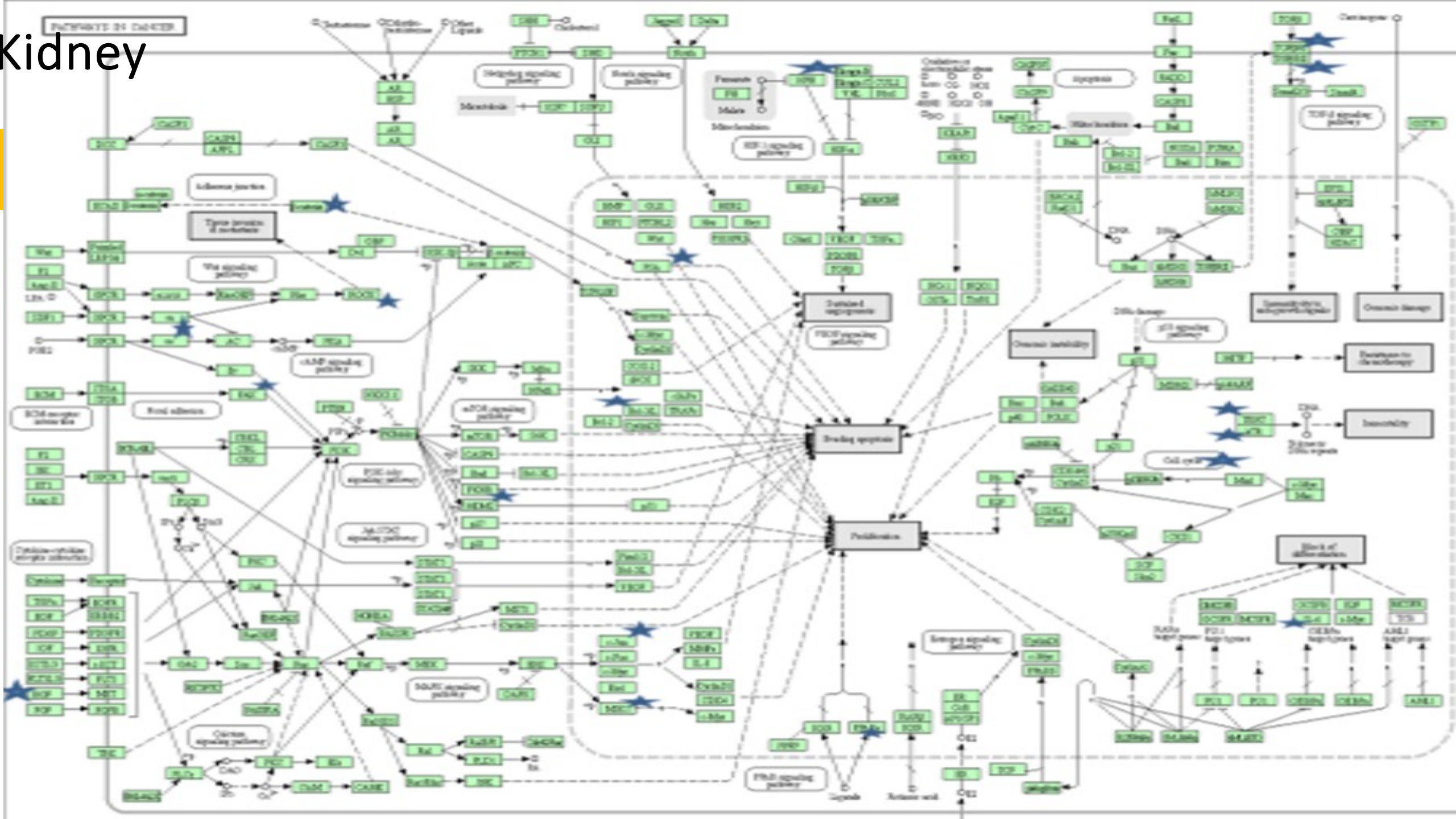
Open Access

Fibroblast growth factor 11 (FGF11) promotes non-small cell lung cancer (NSCLC) progression by regulating hypoxia signaling pathway



David Enrichment Analysis

Kidney



David Enrichment Analysis

Kidney






Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
<input type="checkbox"/>	GOTERM_BP_DIRECT	apoptotic process	RT		453	3.8	5.9E-22	6.7E-18
<input type="checkbox"/>	GOTERM_BP_DIRECT	signal transduction	RT		838	7.0	3.7E-14	2.1E-10
<input type="checkbox"/>	GOTERM_BP_DIRECT	inflammatory response	RT		303	2.5	1.2E-13	4.0E-10
<input type="checkbox"/>	GOTERM_BP_DIRECT	regulation of catalytic activity	RT		280	2.3	1.4E-13	4.0E-10
<input type="checkbox"/>	GOTERM_BP_DIRECT	positive regulation of transcription from RNA polymerase II promoter	RT		792	6.6	5.4E-13	1.1E-9

Lung






Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
<input type="checkbox"/>	GOTERM_BP_DIRECT	protein phosphorylation	RT		376	3.1	3.7E-18	4.2E-14
<input type="checkbox"/>	GOTERM_BP_DIRECT	positive regulation of transcription from RNA polymerase II promoter	RT		814	6.7	3.4E-16	2.0E-12
<input type="checkbox"/>	GOTERM_BP_DIRECT	cell division	RT		289	2.4	9.4E-15	3.6E-11
<input type="checkbox"/>	GOTERM_BP_DIRECT	negative regulation of transcription from RNA polymerase II promoter	RT		658	5.5	8.5E-14	2.4E-10
<input type="checkbox"/>	GOTERM_BP_DIRECT	cell cycle	RT		270	2.2	2.5E-13	5.8E-10

David Enrichment Analysis

Kidney






Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
<input type="checkbox"/>	GOTERM_CC_DIRECT	cytosol	RT		3788	31.7	7.1E-129	1.2E-125
<input type="checkbox"/>	GOTERM_CC_DIRECT	membrane	RT		1876	15.7	8.2E-91	6.8E-88
<input type="checkbox"/>	GOTERM_CC_DIRECT	cytoplasm	RT		3679	30.8	3.3E-79	1.8E-76
<input type="checkbox"/>	GOTERM_CC_DIRECT	nucleoplasm	RT		2682	22.4	1.9E-72	8.0E-70
<input type="checkbox"/>	GOTERM_CC_DIRECT	extracellular exosome	RT		1498	12.5	5.5E-35	1.8E-32

Lung






Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
<input type="checkbox"/>	GOTERM_CC_DIRECT	cytosol	RT		3794	31.5	3.5E-121	5.8E-118
<input type="checkbox"/>	GOTERM_CC_DIRECT	cytoplasm	RT		3711	30.8	8.1E-80	6.8E-77
<input type="checkbox"/>	GOTERM_CC_DIRECT	nucleoplasm	RT		2721	22.6	8.6E-78	4.8E-75
<input type="checkbox"/>	GOTERM_CC_DIRECT	membrane	RT		1843	15.3	1.3E-73	5.4E-71
<input type="checkbox"/>	GOTERM_CC_DIRECT	nucleus	RT		3799	31.5	7.9E-51	2.7E-48

David Enrichment Analysis

Kidney

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
<input type="checkbox"/>	GOTERM_MF_DIRECT	protein binding	RT		8307	69.5	4.4E-175	1.7E-171
<input type="checkbox"/>	GOTERM_MF_DIRECT	identical protein binding	RT		1229	10.3	1.3E-30	2.5E-27
<input type="checkbox"/>	GOTERM_MF_DIRECT	ATP binding	RT		1053	8.8	1.2E-14	1.5E-11
<input type="checkbox"/>	GOTERM_MF_DIRECT	RNA binding	RT		992	8.3	4.3E-12	4.1E-9
<input type="checkbox"/>	GOTERM_MF_DIRECT	macromolecular complex binding	RT		306	2.6	2.5E-10	1.9E-7

Lung

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
<input type="checkbox"/>	GOTERM_MF_DIRECT	protein binding	RT		8340	69.1	2.7E-166	1.0E-162
<input type="checkbox"/>	GOTERM_MF_DIRECT	ATP binding	RT		1103	9.1	1.8E-24	3.3E-21
<input type="checkbox"/>	GOTERM_MF_DIRECT	identical protein binding	RT		1183	9.8	3.3E-18	4.1E-15
<input type="checkbox"/>	GOTERM_MF_DIRECT	RNA binding	RT		1013	8.4	9.5E-15	8.9E-12
<input type="checkbox"/>	GOTERM_MF_DIRECT	protein kinase binding	RT		376	3.1	1.5E-11	1.1E-8

Functional Annotation Chart

[Help and Manual](#)

Current Gene List: list.degs_Lung
Current Background: Homo sapiens
12061 DAVID IDs



 Options

Rerun Using Options

Create Sublist

2 chart records

 [Download File](#)

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
<input type="checkbox"/>	OMIM_DISEASE	Leukemia, acute myeloid, somatic	RT		9	0.1	4.3E-2	1.0E0
<input type="checkbox"/>	OMIM_DISEASE	Breast cancer, somatic	RT		8	0.1	6.9E-2	1.0E0

The related OMIM diseases of our lung DEGs.

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 RL, Miller KD, Jemal A. Cancer statistics, 2020. *CA Cancer J Clin*. 2020 Jan;70(1):7-30. [PubMed] [Reference list]
- [3] Mojic M, Takeda K, Hayakawa Y. The Dark Side of IFN- γ : Its Role in Promoting Cancer Immune Evasion. *Int J Mol Sci*. 2017 Dec 28;19(1):89. doi: 10.3390/ijms19010089. PMID: 29283429; PMCID: PMC5796039.
- [4] [Yang Yang](#), [Bao Liu](#) Exploring and comparing of the gene expression and methylation differences between lung adenocarcinoma and squamous cell carcinoma, *J Cell Physiol*, 2019 Apr;234(4):4454-4459. doi: 10.1002/jcp.27240.
- [5] 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.
- [6] 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. (2010). Lung Squamous Cell Carcinoma mRNA Expression Subtypes Are Reproducible, Clinically Important, and Correspond to Normal Cell Types. *Clinical Cancer Research*, 16(19), 4864–4875. doi:10.1158/1078-0432.CCR-10-0199
- [7] [William Yang](#), [Kenji Yoshigoe](#), [Xiang Qin](#), [Jun S Liu](#), [Jack Y Yang](#), [Andrzej Niemierko](#), [Youping Deng](#), [Yunlong Liu](#), [A Keith Dunker](#), [Zhongxue Chen](#), [Liangjiang Wang](#), [Dong Xu](#), [Hamid R Arabnia](#), [Weida Tong](#) & [Mary Qu Yang](#), Identification of genes and pathways involved in kidney renal clear cell carcinoma, [BMC Bioinformatics](#) volume 15, Article number: S2 (2014)
- [8] Kai Wang, Xiao Qu, Shaorui Liu, Xudong Yang, Fenglong Bie, Yu Wang, Cuicui Huang, Jiajun Du, Identification of aberrantly expressed F-box proteins in squamous-cell lung carcinoma, *Journal of Cancer Research and Clinical Oncology* (2018) 144:1509–1521
- [9] Yijun Qi, Yue Zhang, Zhiqiang Peng, Lei Wang, Kaizhen Wang, Duiping Feng, Junqi He, Junfang Zheng, 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.* Vol 22, No 2, 2018 pp. 1224-1235
- [10] Anbang He, Shiming He, Cong Huang, Zhicong Chen, Yucai Wu, Yanqing Gong, Xuesong Li, Liqun Zhou, MTDH promotes metastasis of clear cell renal cell carcinoma by activating SND1-mediated ERK signaling and epithelial-mesenchymal transition, *AGING* 2020, Vol. 12, No. 2



Thank you