DIFFERENTIAL ANALYSES OF GENE EXPRESSION

DILARA ISIKLI & RICCARDO CECCARONI
GROUP 08

CONTENTS

LIST OF FIGURES

ABSTRACT

Lung adenocarcinoma (LUAD) is the leading cause of cancer-related death worldwide. The main obstacle to early diagnosis or monitoring of patients at high risk of poor survival has been the lack of essential predictive biomarkers.

RNA-sequencing was performed on LUAD affected tissue and paired adjacent to non-cancerous tissue samples. The Cancer Genome Atlas project-LUAD dataset was used to obtain an intersection of differential expressed genes.

In this study 494 candidate genes is identified (237 up regulated and 257 down regulated genes) with $|\text{fold change}| \geqslant 2.5$ and $p \leqslant 0.05$ in DEGs analysis.

Using Gene co-expression network and Differential Gene co-expression network the genes that characterize lung cancer is identified.

1 INTRODUCTION

Lung cancer is the leading cause of cancer-related deaths globally [1]. LUAD accounts for approximately 40% of all cases [2]. Over the past several decades, in spite of the current multimodal therapy, the survival time of LUAD patients has shown marginal improvement only. LUAD recurrence and metastasis are common, even with the tumor diagnosed at an early stage. [3] It is necessary to identify novel biomarkers and therapeutic targets for treatment of LUAD. With the development of high-throughput technology, gene expression profiles have been broadly used to identify more novel biomarkers. RNA-sequencing (RNA-seq) technology is an efficient high-throughput sequencing tool to measure transcripts, identify new transcriptional units and discover differentially expressed genes (DEGs) among samples. RNA-seq, usually together with bioinformatics methods, has been broadly used in cancer research. For example, recent studies have found several key genes in lung cancer using RNA-seq and bioinformatics methods. [4] [5]

In this study 494 candidate genes is identified (237 up regulated and 257 down regulated genes) with $|\text{fold change}| \geqslant 2.5$ and $p \leqslant 0.05$ in DEGs analysis.

Using Gene co-expression network we identify 5 genes with high degree present only in the cancer condition (BUB1B, KIF4A, DLGAP5, SKA3, EXO1) and 8 genes with high degree present only in the normal condition (GTSE1, NDC80, CDC6, TOP2A, NUSAP1 , CEP55, CDCA5, SKA1). Finally, using Differential Gene co-expression network we identified 10 genes: ZMYND10, NGEF, NEK2, CDHR3, PEBP4, PLPP2, SFTPC, SFTPB, BUB1, MAP7D2.

2 MATERIALS AND METHODS

All code and key data files for this analyses are available in the GitHub folder [1].

2.1 Data

The research data source is https://portal.gdc.cancer.gov/ . The TCGA-LUAD project is selected in the GDC data portal. The data are filtered with Transcriptome Profiling as data category, Gene Expression Quantification as data type and HTSeq-FPKM as workflow type. Finally, only patients for whom cancer and normal tissue files are available are selected. A data set with 57 patients and 17224 genes is obtained for both normal condition (dataN) and cancer condition (dataC).

2.2 Differentially Expressed Genes

A first criterion to find differentially expressed genes can be to identify the genes whose expression in the two groups (normal and cancer) of considered samples varies by a certain proportion. The fold-change is calculated using the following formula:

$$FC = \frac{\log_2(dataN)}{\log_2(dataC)}$$

The values obtained are shown on the following histogram (Figure 1).

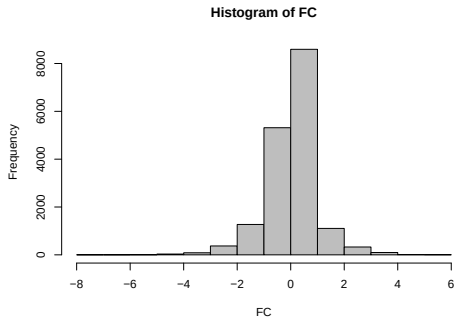


Figure 1

1 https://github.com/ceccaroni1884368/DE_TCGA-LUAD

Another criterion to find differentially expressed genes is to use Student's t test for two conditions. So t-test is used to calculate the p-value. The "fdr" is applied method for correction multiple comparison. The values obtained are shown on the following histogram (Figure 2).

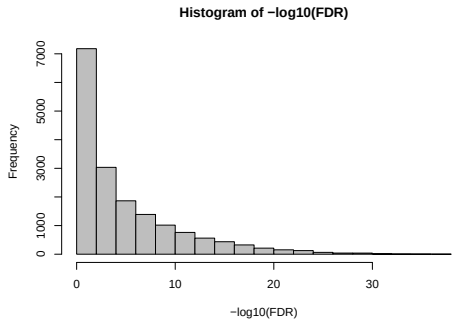


Figure 2

We have selected | fold change | $\geqslant 2.5$ and $\texttt{fdr} \leqslant 0.05$ as threshold values. The result is the volcano plot in Figure 3.

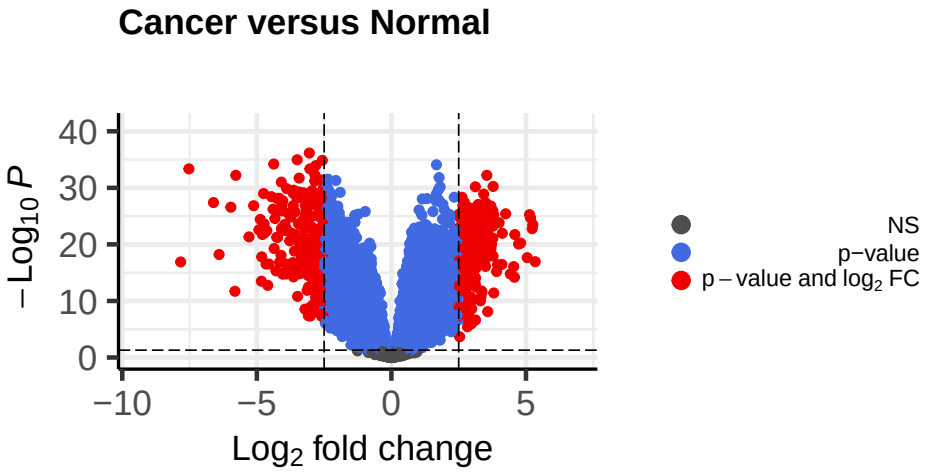


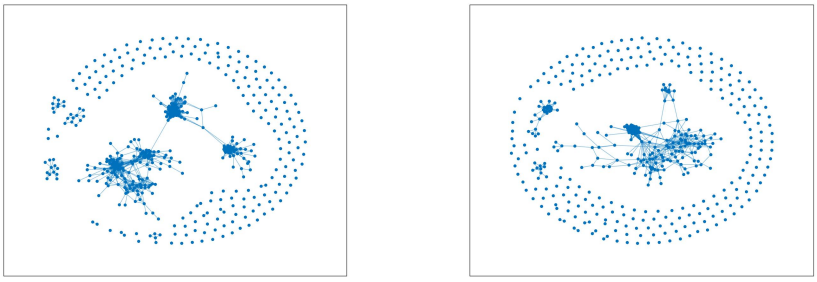
Figure 3

In the end 494 genes (237 upregulated and 257 downregulated genes) were found.

2.3 Co-expression networks

A gene co-expression network graph is calculated to show is each node corresponds to a gene, and a pair of nodes is connected with an edge if there is a significant co-expression relationship between them.

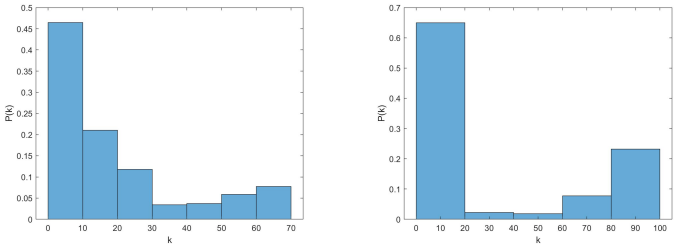
To calculation, significance threshold is selected, $\log_2(\cdot)$ operator is applied to the data then correlation is calculated.



(a) Normal cells **(b)** Cancer cells

Figure 4: Co-expression network

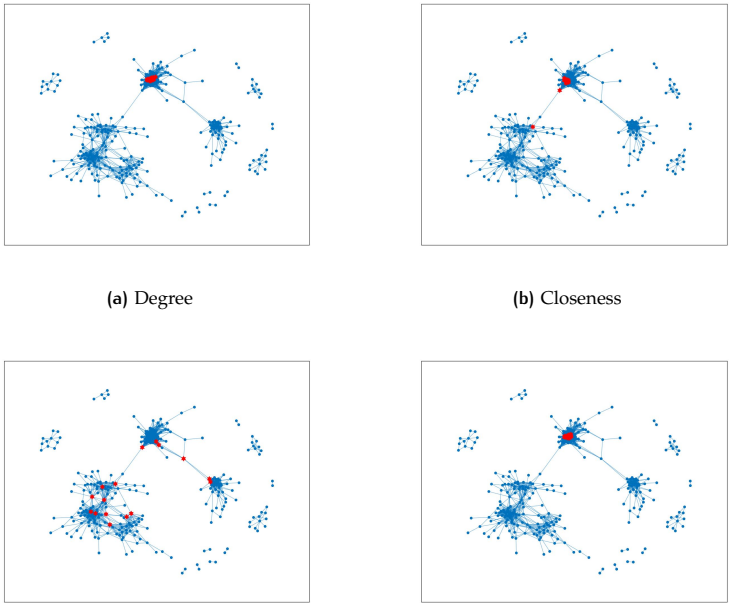
To show degree distribution of network, degree distribution of data is calculated and printed as histogram Figure 5.



(a) Normal cells **(b)** Cancer cells

Figure 5: Degree distribution of co-expression network in normal and cancer cells

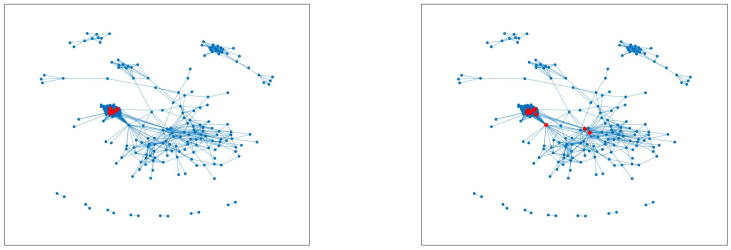
To understand the hubs characterizing, graph centrality measurements is used with highest degree values (5% of the nodes).



(a) Degree **(b)** Closeness

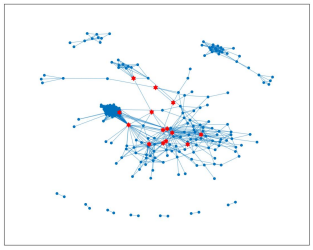
(c) Betweenness **(d)** Eigenvector

Figure 6: Centrality Measures in Normal Cells

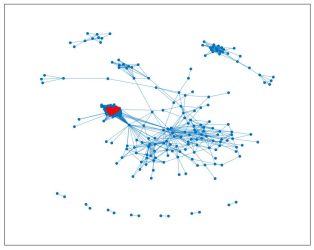


(a) Degree

(b) Closeness



(c) Betweenness



(d) Eigenvector

Figure 7: Centrality Measures in Cancer Cells

2.4 Differential Co-expressed Network

In this part, instead of establishing that co-expression is significant in one condition and not in the other, directly it is tested if the change in co-expression is significant using differential networks: they encode changes in the connections among nodes between the conditions or states.

To calculate the differential correlations, the variance of sample correlation coefficients is stabilized in each condition applying the following Fisher z- transformation:

$$z_{1or2} = \frac{1}{2}\log\left(\frac{1+\rho_{1or2}}{1-\rho_{1or2}}\right)$$

then z-scores is computed to evaluate the correlation:

$$Z = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}}$$

where n_1 and n_2 represent the sample size for each of the conditions. Finally, $|Z| > 5$ as threshold is set; the Figure 8 is showed up.

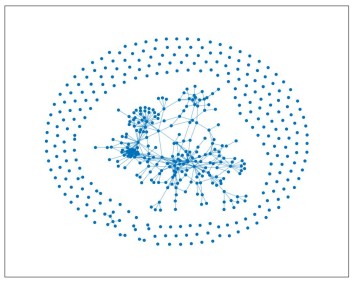


Figure 8: Differentially Co-expression network

The degree distribution is calculated and printed as histogram Figure 9.

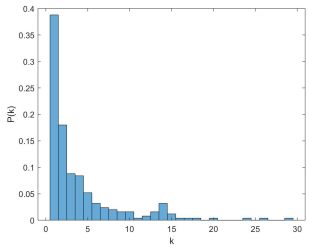
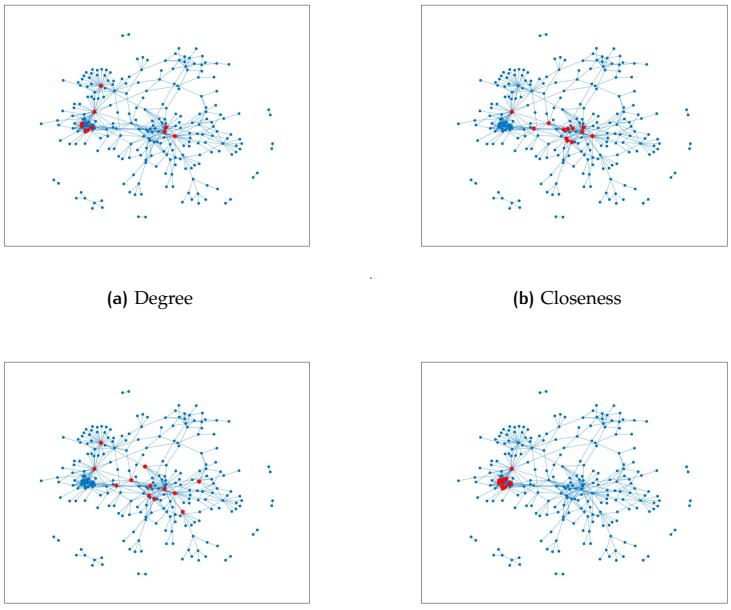


Figure 9: Degree distribution of differentially co-expression network

As before, the graphs of the centrality measures are displayed in the Figure 10.



(a) Degree **(b)** Closeness

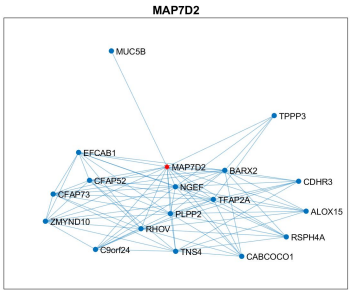
(c) Betweenness **(d)** Eigenvector

Figure 10: Centrality Measures

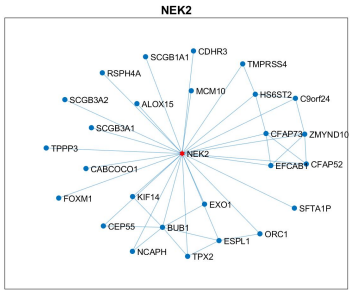
In the next paragraph all the sub-graphs of the nodes with the highest degree value in differential co-expression network will be illustrated.

Subnetwork plot of the most relevant genes

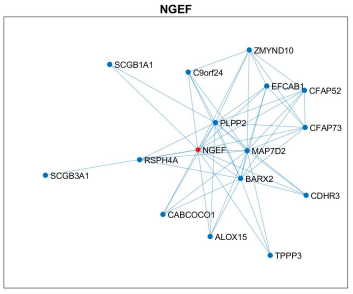
(a) **(b)**


MAP7D2
MUC5B
TPPP3
EFCAB1
MAP7D2
BARX2
CFAP52
NGEF
CDHR3
CFAP73
TFAP2A
ALOX15
PLPP2
ZMYND10
RHOV
RSPH4A
C9orf24
TNS4
CABCOCO1

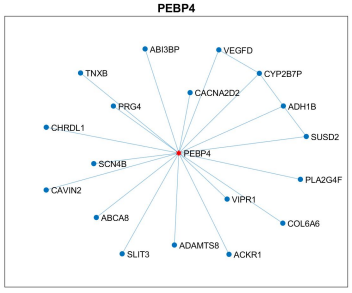
(c)


NEK2
SCGB1A1
CDHR3
TMPRSS4
RSPH4A
MCM10
HS6ST2
C9orf24
SCGB3A2
ALOX15
SCGB3A1
CFAP73
ZMYND10
TPPP3
NEK2
CABCOCO1
EFCAB1
CFAP52
FOXM1
KIF14
EXO1
SFTA1P
CEP55
BUB1
ORC1
ESPL1
NCAPH
TPX2

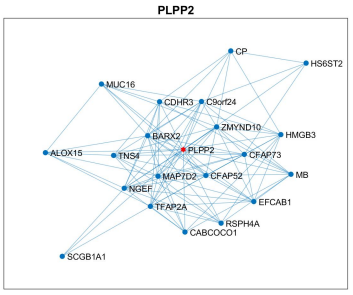
(d)


NGEF
ZMYND10
SCGB1A1
C9orf24
EFCAB1
CFAP52
PLPP2
CFAP73
NGEF
MAP7D2
RSPH4A
SCGB3A1
BARX2
CDHR3
CABCOCO1
ALOX15
TPPP3

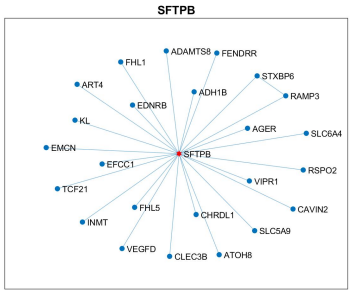
(e)


PEBP4
ABI3BP
VEGFD
TNXB
CYP2B7P
CACNA2D2
PRG4
ADH1B
CHRDL1
SUSD2
PEBP4
SCN4B
PLA2G4F
CAVIN2
VIPR1
ABCA8
COL6A6
SLIT3
ADAMTS8
ACKR1

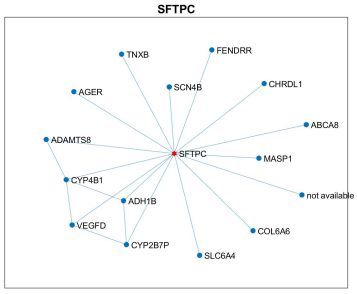
(f)


PLPP2
CP
HS6ST2
MUC16
CDHR3
C9orf24
ZMYND10
BARX2
HMGB3
ALOX15
TNS4
PLPP2
CFAP73
MAP7D2
CFAP52
MB
NGEF
EFCAB1
TFAP2A
RSPH4A
CABCOCO1
SCGB1A1

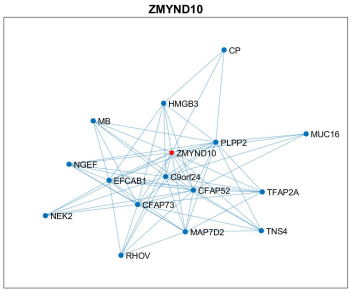
(g)


SFTPB
ADAMTS8
FENDRR
FHL1
STXBP6
ART4
ADH1B
RAMP3
EDNRB
KL
AGER
SLC6A4
EMCN
SFTPB
EFCC1
RSPO2
VIPR1
TCF21
FHL5
CAVIN2
CHRDL1
INMT
SLC5A9
VEGFD
CLEC3B
ATOH8

(h)


SFTPC
TNXB
FENDRR
AGER
SCN4B
CHRDL1
ABCA8
ADAMTS8
SFTPC
MASP1
CYP4B1
not available
ADH1B
VEGFD
COL6A6
CYP2B7P
SLC6A4

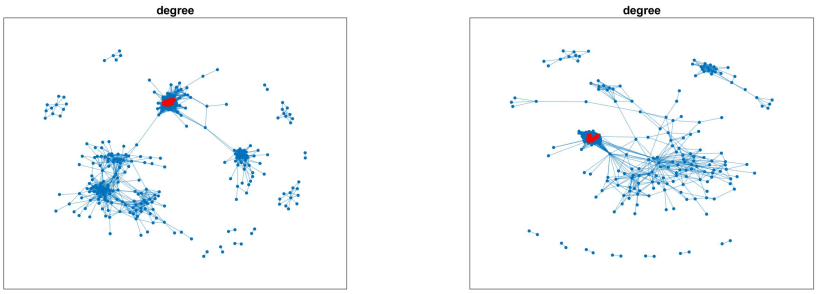
(i)


ZMYND10
CP
HMGB3
MB
MUC16
PLPP2
ZMYND10
NGEF
EFCAB1
C9orf24
CFAP52
TFAP2A
CFAP73
NEK2
MAP7D2
TNS4
RHOV

(j)

3 RESULTS AND DISCUSSION

3.1 Co-expression networks



(a) Normal cells **(b)** Cancer cells

Figure 12: Co-expression network compare hub sets

Identify the hubs characterizing cancer tissue

For cancer cells, TPX2, BUB1B, KIF4A, HJURP, NCAPG, DLGAP5, MELK, SKA3, CP2L, EXO1 are found with considering only degrees which bigger than 95%. To making hub identification, first of all values of degrees, betweeness, closeness, eigenvetor which belong to each gene are used. The details are seen in the table below (Table 1):

Gene	Degree	Betweeness	Closeness	Eigenvector
TPX2	89	200	0.00120	0.00979
BUB1B	90	194	0.00113	0.00997
KIF4A	90	213	0.00113	0.01012
HJURP	90	200	0.00116	0.00974
NCAPG	89	189	0.00120	0.00978
DLGAP5	89	184	0.00120	0.00981
MELK	89	15.9	0.00113	0.00982
SKA3	89	181	0.00116	0.00982
CKAP2L	90	13.4	0.00121	0.00986
EXO1	89	189	0.00120	0.00978

Table 1: Co-expression network compare hub sets

Then these are searched in the lists of betweenness (over to 95%), closeness (over to 95 percent), eigenvetor (over to 95%) and degree (over to 99%). The details are seen in the table below (Table 2):

Gene	Degree	Betweeness 95%	Closeness 95%	Eigenvector 95%	Degree 99%
TPX2	89	NO	YES	YES	NO
BUB1B	90	NO	YES	YES	YES
KIF4A	90	NO	YES	YES	YES
HJURP	90	NO	YES	YES	NO
NCAPG	89	NO	YES	YES	NO
DLGAP5	89	NO	YES	YES	NO
MELK	89	NO	NO	YES	NO
SKA3	89	NO	YES	YES	NO
CKAP2L	90	NO	YES	YES	YES
EXO1	89	NO	YES	YES	NO

Table 2: Co-expression network compare hub sets

Compare hubs sets related to cancer and normal conditions

TPX2, BUB1B, KIF4A, HJURP, NCAPG, DLGAP5, MELK, SKA3, CKAP2L, EXO1 are found with considering only degrees which bigger than 95% in cancer condition.

TPX2, GTSE1, NDC80, CDC6, TOP2A, NUSAP1, CEP55, CDCA5, SKA1, HJURP, NCAPG, MELK, CKAP2L are found with considering only degrees which bigger than 95% in normal condition.

To compare, these are searched in the lists of betweenness (over to 95%), closeness (over to 95%), eigenvetor (over to 95%) and degree(over to 95%) in both conditons. Also as is seen, HJURP, NCAPG, MELK, CKAP2L and TPX2 are exist in both condition. The details are seen in the table below (Table 3):

Gene	Degree 95% Normal	Degree 95% Cancer	Betweeness 95% Normal	Betweeness 95% Cancer	Closeness 95% Normal	Closeness 95% Cancer	Eigenvector 95% Normal	Eigenvector 95% Cancer
TPX2 *	YES	YES	NO	NO	NO	YES	NO	YES
BUB1B	NO	YES	NO	NO	NO	YES	NO	YES
KIF4A	NO	YES	NO	NO	NO	YES	NO	YES
HJURP *	YES	YES	NO	NO	NO	YES	NO	YES
NCAPG *	YES	YES	NO	NO	NO	YES	NO	YES
DLGAP5	NO	YES	NO	NO	NO	YES	NO	YES
MELK *	YES	YES	NO	NO	NO	NO	NO	YES
SKA3	NO	YES	NO	NO	NO	YES	NO	YES
CKAP2L *	YES	YES	NO	NO	NO	YES	NO	YES
EXO1	NO	YES	NO	NO	NO	YES	NO	YES
GTSE1	YES	NO	NO	NO	NO	NO	NO	NO
NDC80	YES	NO	NO	NO	NO	NO	NO	NO
CDC6	YES	NO	NO	NO	NO	NO	NO	NO
TOP2A	YES	NO	NO	NO	NO	NO	NO	NO
NUSAP1	YES	NO	NO	NO	NO	NO	NO	NO
CEP55	YES	NO	NO	NO	NO	NO	NO	NO
CDCA5	YES	NO	NO	NO	NO	NO	NO	NO
SKA1	YES	NO	NO	NO	NO	NO	NO	YES

Table 3: Co-expression network compare hub sets

3.2 Differential Co-expression networks

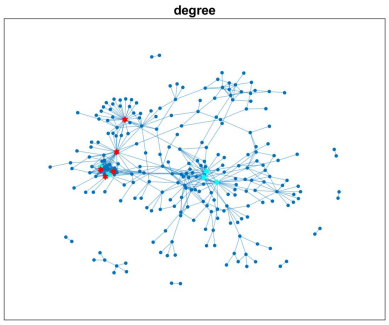


Figure 13: Differentially Co-expression network compare hub sets

Compare the identified hubs set with those obtained in analysis of co-expression networks

To compare the identified hubs set with those obtained in analysis of co-expression networks, first of all ZMYND10, NGEF, NEK2, CDHR3, PEBP4, PLPP2, SFTPC, SFTPB, BUB1, MAP7D2 are found with considering only degrees which bigger than 95%.

Then to detail, these are searched in the lists of betweenness (over to 95%), closeness (over to 95%), eigenvetor (over to 95%) and degree(over to 95%). The details are seen in the table below (Table 4):

Gene	Degree 95%	Betweeness 95%	Closeness 95%	Eigenvector 95%
ZMYND10	YES	NO	NO	YES
NGEF	YES	NO	NO	YES
NEK2	YES	YES	YES	YES
CDHR3	YES	NO	NO	YES
PEBP4	YES	YES	YES	YES
PLPP2	YES	NO	NO	YES
SFTPC	YES	NO	YES	NO
SFTPB	YES	YES	YES	NO
BUB1	YES	YES	NO	NO
MAP7D2	YES	NO	NO	YES

Table 4: Differential Co-expression network compare hub sets

As is seen, there is no any common element between degree list of this analysis and degree list of analysis of co-expression networks.

Also, degree list of this analysis were searched in the lists of betweenness (over to 95%), closeness (over to 95%), eigenvetor (over to 95%) and degree(over to 95%) of analysis of co-expression networks. But none of them was found.

On the other hand, degree list of analysis of co-expression networks were searched in the lists of betweenness (over to 95%), closeness (over to 95%), eigenvetor (over to 95%) and degree (over to 95%) of this analysis. And none of them was found.

Related works

In other studies, ZMYND10 was labeled as a potential tumor suppressor gene in multiple tumor types. It is frequently inactivated or downregulated via genetic or epigenetic changes in many solid tumors, such as lung cancer [6].

NEK2 is one of the best proliferation markers in the prognosis of non-small cell lung cancer (NSCLC)[7][8]. Using a previously validated antibody, NEK2 staining was observed to be concentrated in the nucleus of breast and lung cancer cells [9]. NEK2 is a promising predictor of poor prognosis in cancer because its expression is highly correlated with rapid relapse and poor outcome in multiple cancer types[10].

The overexpression of PEBP4 increases the phosphorylation levels of Akt and mTOR in lung cancer cells. The PI3K/Akt/mTOR signaling axis may be a key molecular pathway via which PEBP4 promotes the proliferation and invasion of non-small cell lung cancer (NSCLC) cells; also, it may serve as a potential therapeutic target[11].

REFERENCES

[1] Rebecca Siegel, Kimberly Miller, and Ahmedin Jemal. Cancer statistics, 2020. *CA: A Cancer Journal for Clinicians*, 70, 01 2020.

[2] David Ettinger, Douglas Wood, Wallace Akerley, Lyudmila Bazhenova, Hossein Borghaei, David Camidge, Richard Cheney, Lucian Chirieac, Thomas D'Amico, Todd Demmy, Thomas Dilling, Ramaswamy Govindan, Frederic Grannis, Leora Horn, Thierry Jahan, Ritsuko Komaki, Mark Kris, Lee Krug, Rudy Lackner, and Miranda Hughes. Non–small cell lung cancer, version 1.2015. *Journal of the National Comprehensive Cancer Network*, 12:1738–1761, 12 2014.

[3] Hyeon-Kyoung Koo, Sang-Man Jin, Chang-Hoon Lee, Hyo-Jeong Lim, Jae-Joon Yim, Young Kim, Seok-Chul Yang, Chul-Gyu Yoo, Sung Han, Joo Kim, Young-Soo Shim, and Young Kim. Factors associated with recurrence in patients with curatively resected stage i-ii lung cancer. *Lung cancer (Amsterdam, Netherlands)*, 73:222–9, 12 2010.

[4] Linlin Xue, Li Xie, Xingguo Song, and Xianrang Song. Identification of potential tumor-educated platelets rna biomarkers in non-small-cell lung cancer by integrated bioinformatical analysis. *Journal of Clinical Laboratory Analysis*, 32, 02 2018.

[5] Shicheng Li, Xiao Sun, Shuncheng Miao, Jia Liu, and Wenjie Jiao. Differential protein-coding gene and long noncoding rna expression in smoking-related lung squamous cell carcinoma. *Thoracic Cancer*, 8, 09 2017.

[6] Yan Wang, Liangying Dan, Qianqian Li, Lili Li, Lan Zhong, Shao Bianfei, Fang Yu, Sanxiu He, Shaorong Tian, Jin He, Qian Xiao, Thomas Putti, Shaquila He, Yixiao Feng, Yong Lin, and Tingxiu Xiang. Zmynd10, an epigenetically regulated tumor suppressor, exerts tumor-suppressive functions via mir145-5p/nedd9 axis in breast cancer. *Clinical Epigenetics*, 11, 12 2019.

[7] Xinwen Zhong, Xiaojiao Guan, Wenke Liu, and Lin Zhang. Aberrant expression of nek2 and its clinical significance in non-small cell lung cancer. *Oncology letters*, 8:1470–1476, 10 2014.

[8] Xinwen Zhong, Xiaojiao Guan, Dong Qianze, Shize Yang, Wenke Liu, and Lin Zhang. Examining nek2 as a better proliferation marker in non-small cell lung cancer prognosis. *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine*, 35, 04 2014.

[9] Chiara Naro, Federica Barbagallo, Paolo Chieffi, Cyril Bourgeois, Maria Paola Paronetto, and Claudio Sette. The centrosomal kinase nek2 is a novel splicing factor kinase involved in cell survival. *Nucleic acids research*, 42, 12 2013.

[10] TOSHIO KOKURYO, YUKIHIRO YOKOYAMA, JUNPEI YAMAGUCHI, NOBUYUKI TSUNODA, TOMOKI EBATA, and Masato Nagino. Nek2 is an effective target for cancer therapy with potential to induce regression of multiple human malignancies. *Anticancer Research*, 39:2251–2258, 05 2019.

[11] Guiping Yu, Bin Huang, Guoqiang Chen, and Yedong Mi. Phosphatidylethanolamine-binding protein 4 promotes lung cancer cells proliferation and invasion via pi3k/akt/mtor axis. *Journal of thoracic disease*, 7:1806–1816, 12 2015.