DIFFERENTIAL ANALYSES OF GENE EXPRESSION

JOHN SMITH & JAMES SMITH
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 noncancerous tissue samples. The Cancer Genome Atlas project-LUAD dataset was used to obtain an intersection of differential expressed genes.

In our stydy we identified 494 candidate genes (237 upregulated and 257 downregulated genes) with $|\text{fold change}| \geqslant 2.5$ and $p \leqslant 0.05$.

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]

2 MATERIALS AND METHODS

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

2.1 Data

The data used in our research come from `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. We calculated the fold-change using the following formula:

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

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

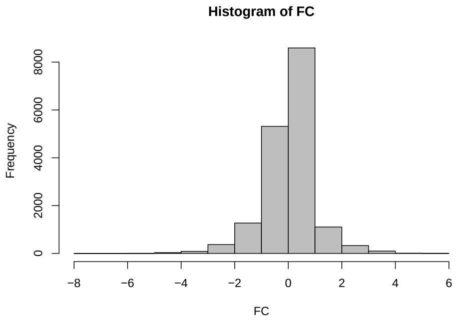


Figure 1

[1] http://www.github.com

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

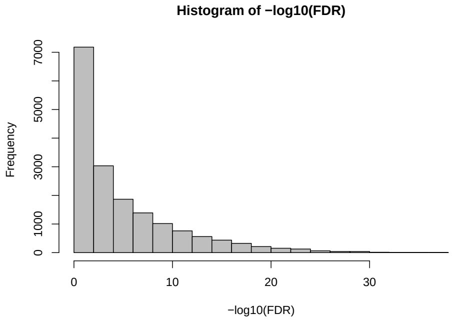


Figure 2

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.

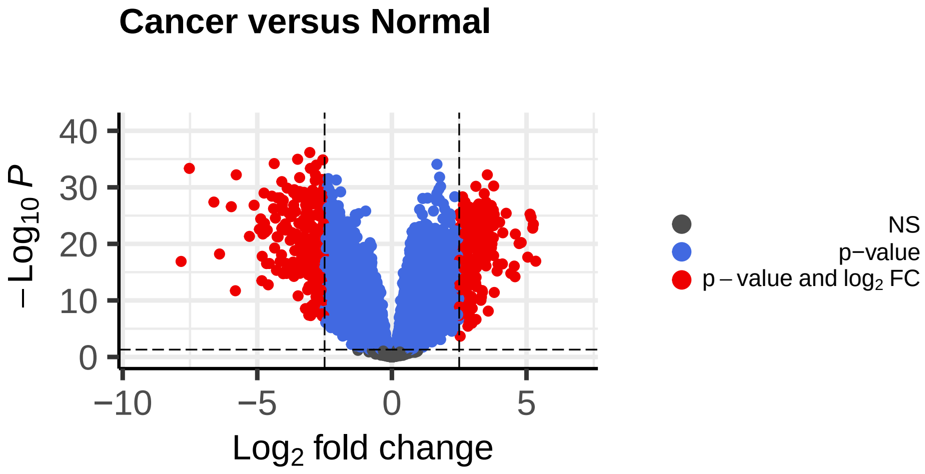
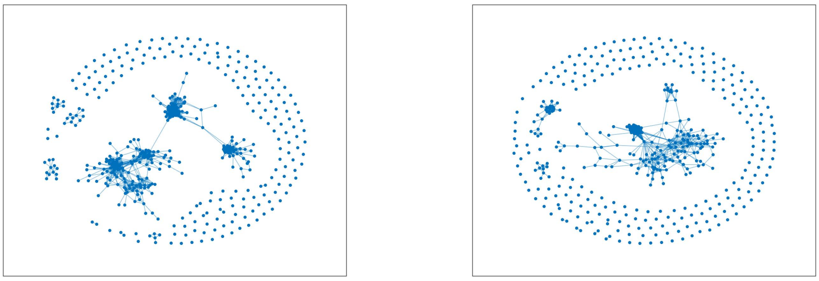


Figure 3

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

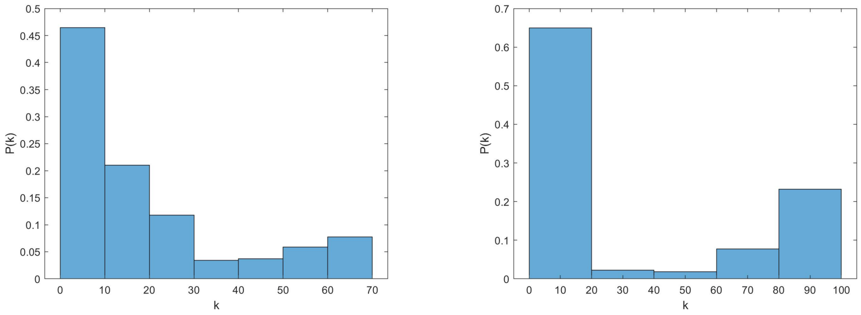
2.3 Co-expression networks



(a) Normal cells

(b) Cancer cells

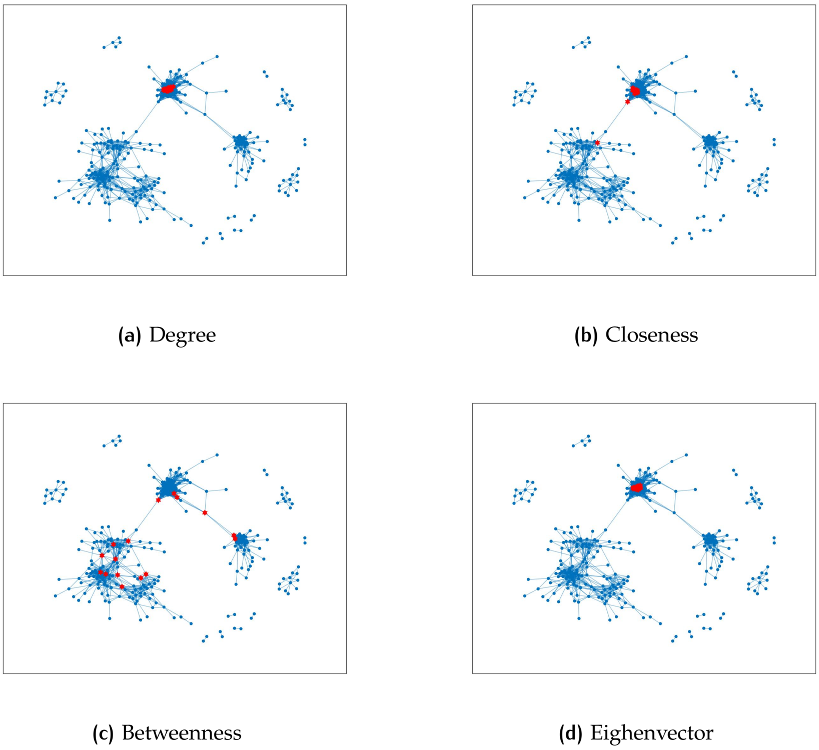
Figure 4: Co-expression network



(a) Normal cells

(b) Cancer cells

Figure 5: Co-expression network in normal and cancer cells



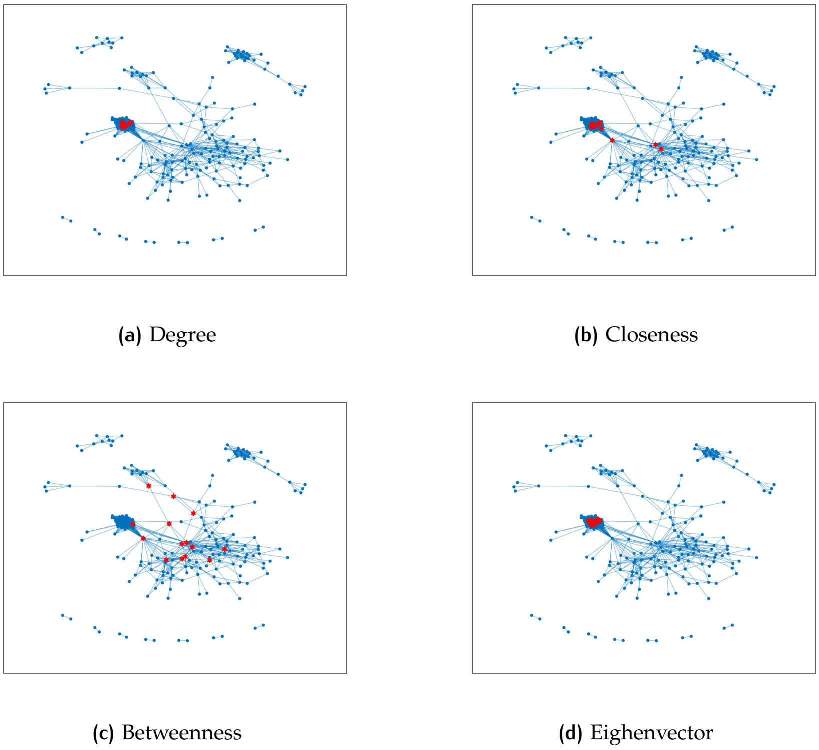
(a) Degree

(b) Closeness

(c) Betweenness

(d) Eighenvector

Figure 6: Centrality Measures in Normal Cells



(a) Degree

(b) Closeness

(c) Betweenness

(d) Eighenvector

Figure 7: Centrality Measures in Cancer Cells

2.4 Differential Co-expressed Network

Instead of establishing that co-expression is significant in one condition and not in the other, we are now going to test directly 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, first we have stabilized the variance of sample correlation coefficients in each condition applying the following Fisher z-transformation:

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

then we compute z-scores 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 we set $|Z| > 5$ as threshold; we get the graph of the Figure 8.

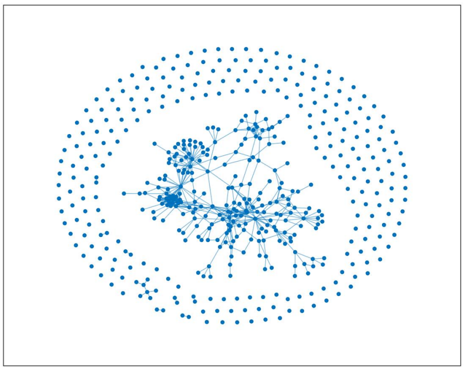


Figure 8: Differentially Co-expression network

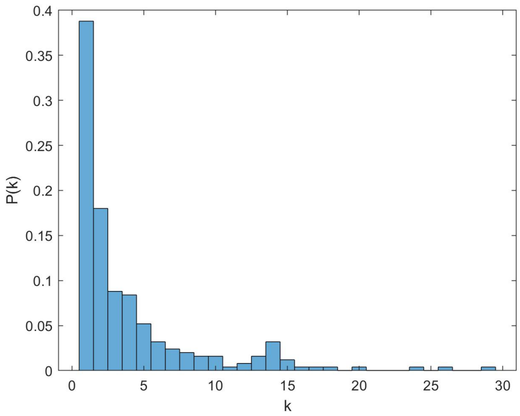
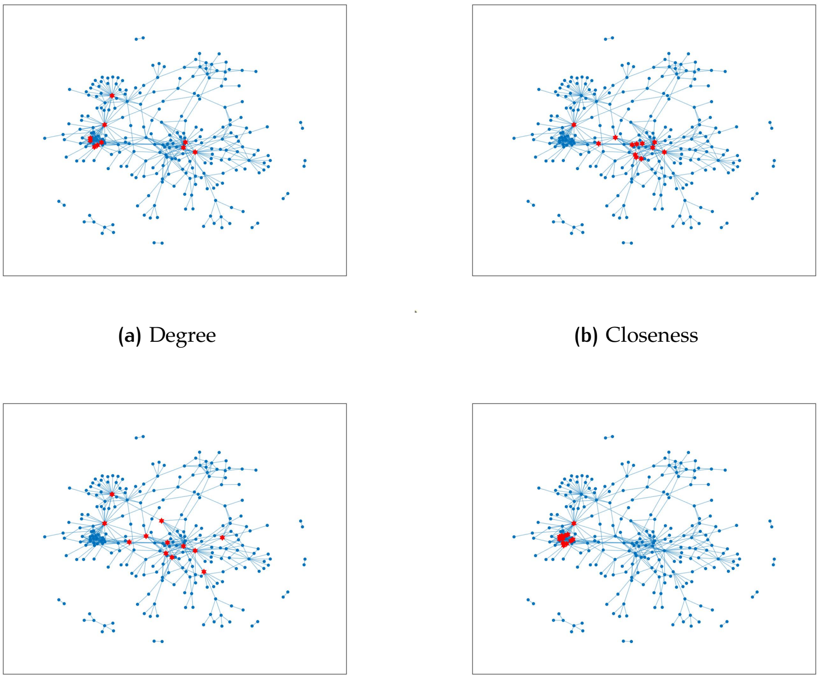


Figure 9: Differentially Co-expression network Degree Distribution



(a) Degree

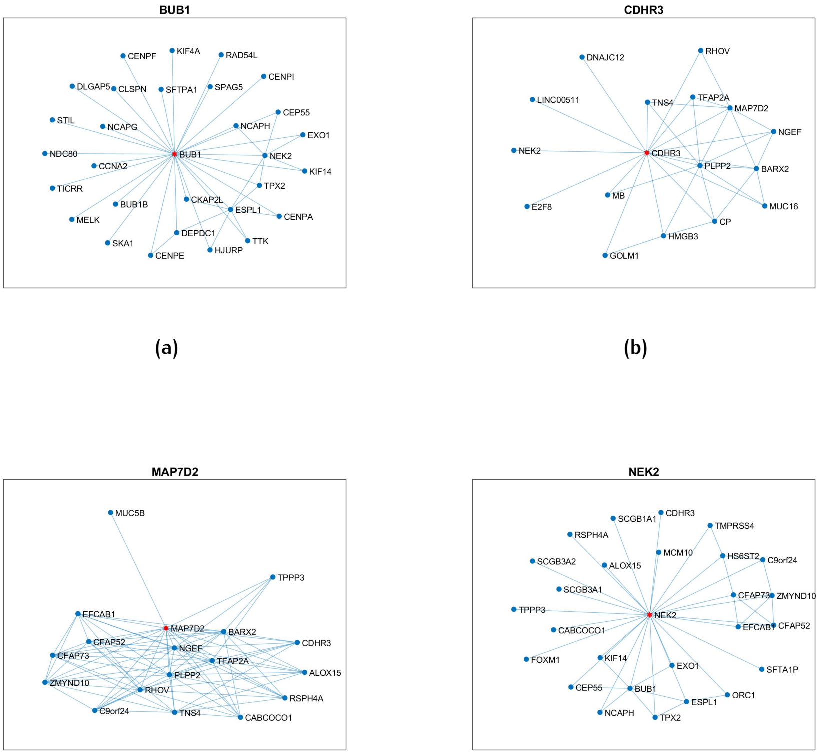
(b) Closeness

(c) Betweenness

(d) Eighenvector

Figure 10: Centrality Measures

2.4.1 Subnetwork plot of the most relevant genes

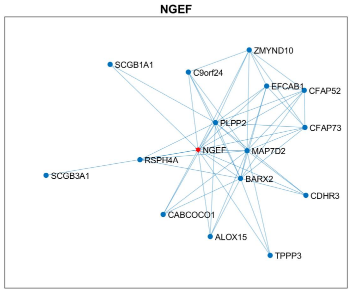


(a)

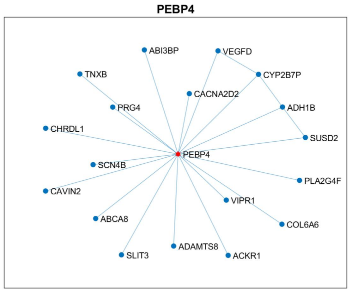
(b)

(c)

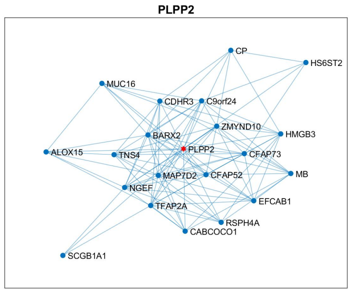
(d)


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

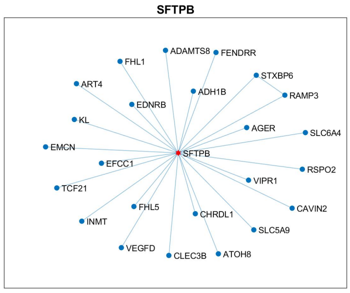
(e)


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

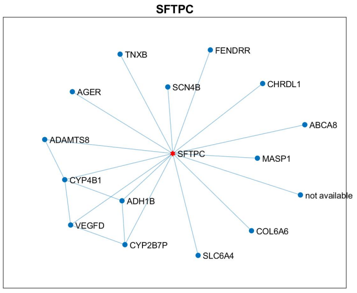
(f)


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

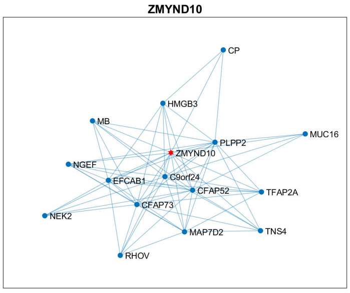
(g)


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

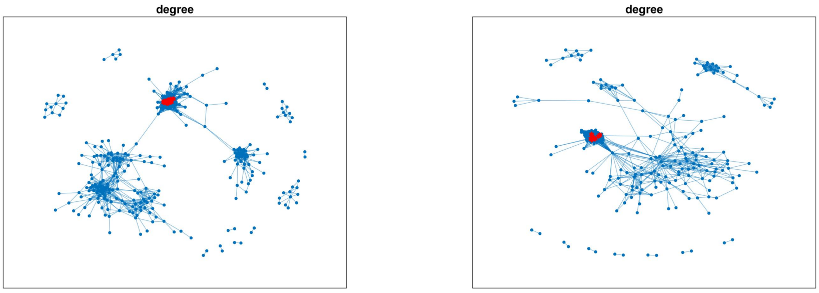
(i)


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

(j)

3 RESULTS AND DISCUSSION

Co-expression networks



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

Figure 12: Co-expression network compare hub sets

Gene	Degree	Betweeness	Closeness	Eigenvector	Betweeness 95%	Closeness 95%	Eigenvector 95%	Degree 99%
TPX2	89	200	0.00120	0.00979	NO	YES	YES	NO
BUB1B	90	194	0.00113	0.00997	NO	YES	YES	YES
KIF4A	90	213	0.00113	0.01012	NO	YES	YES	YES
HJURP	90	200	0.00116	0.00974	NO	YES	YES	NO
NCAPG	89	189	0.00120	0.00978	NO	YES	YES	NO
DLGAP5	89	184	0.00120	0.00981	NO	YES	YES	NO
MELK	89	15.9	0.00113	0.00982	NO	NO	YES	NO
SKA3	89	181	0.00116	0.00982	NO	YES	YES	NO
CKAP2L	90	13.4	0.00121	0.00986	NO	YES	YES	YES
EXO1	89	189	0.00120	0.00978	NO	YES	YES	NO

Table 1: Co-expression network compare hub sets

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 2: Co-expression network compare hub sets

Differential Co-expression networks

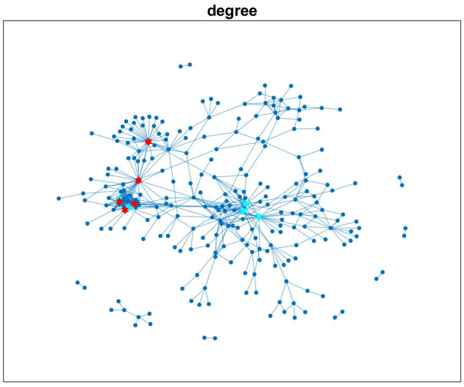


Figure 13: Differentially Co-expression network compare hub sets

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 3: Differential Co-expression network compare hub sets

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.