DIFFERENTIAL ANALYSES OF GENE EXPRESSION

JOHN SMITH & JAMES SMITH

GROUP NN

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 down-regulated 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
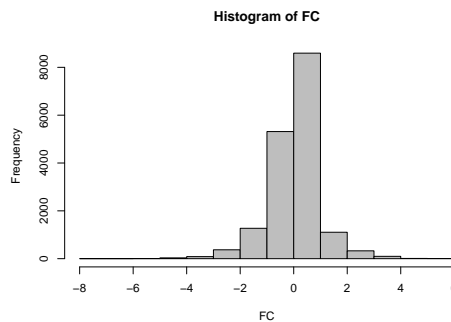
Another criterion to find differentially expressed genes is to use Student's t test for two conditions. So we used a t-test to calculate the p-value. We applied the "fdr" method for correction multiple comparison. The values obtained are shown on the following histogram (Figure 2).
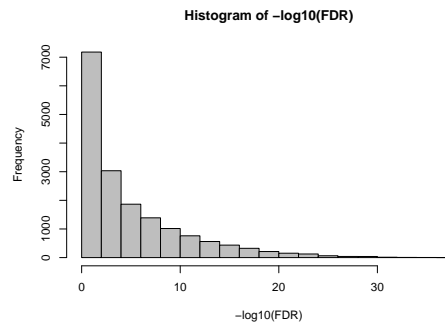
Histogram of −log10(FDR)

Figure 2

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

Cancer versus Normal
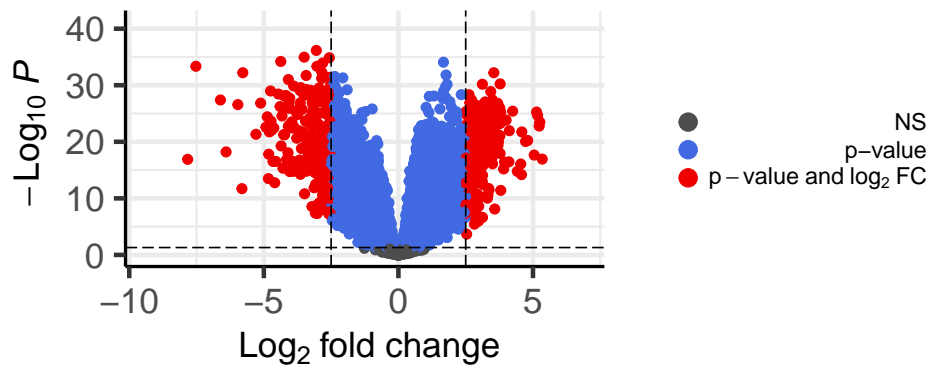
Figure 3

In the end 494 genes (237 upregulated and 257 downregulated genes) were found.
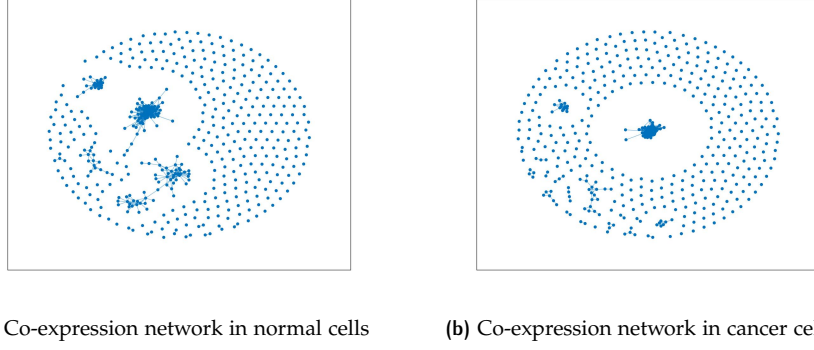
2.3 Co-expression networks

(a) Co-expression network in normal cells   **(b)** Co-expression network in cancer cells

Figure 4: Co-expression network in normal and 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:

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

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 5.
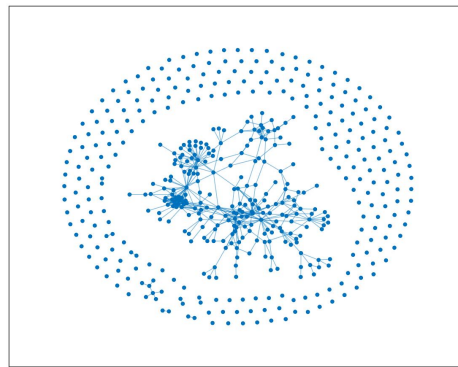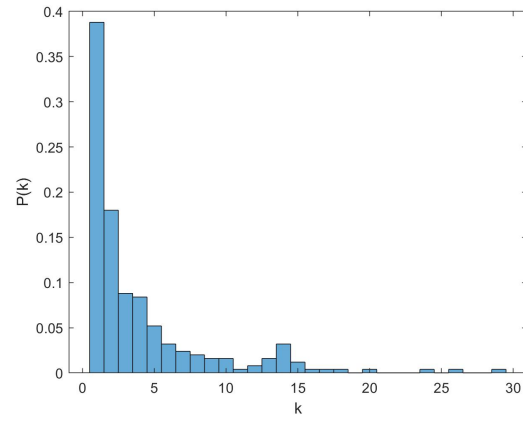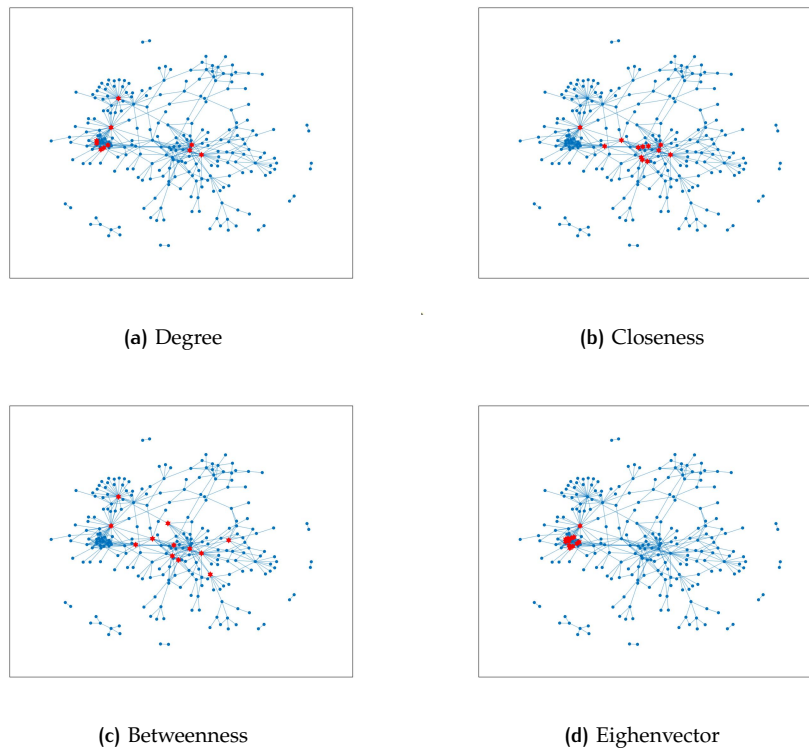

Figure 5: Differentially Co-expression network

Figure 6: Differentially Co-expression network Degree Distribution

(a) Degree

(b) Closeness

(c) Betweenness

(d) Eighenvector

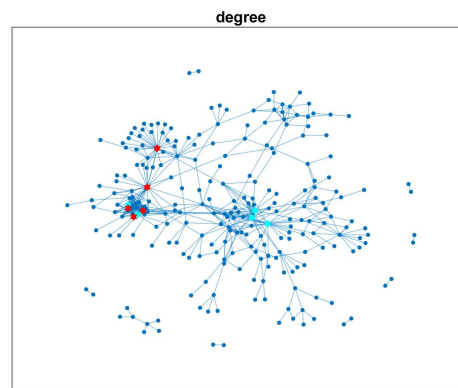**Figure 7:** Centrality Measures

3 RESULTS AND DISCUSSION

Figure 8: Differentially 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.