An analysis for identifying the subtype-specific genes using TCGA lung cancer data set

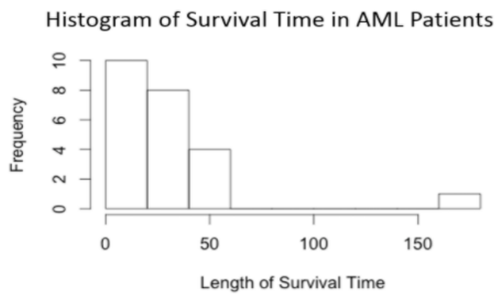
Jaeho Jeong[1] and Kipoong Kim[1]
[1]Department of Statistics, Pusan National University

1. INTRODUCTION

- Lung cancer is traditionally classified as non-small-cell lung cancer (NSCLC) and small-cell lung cancer (SCLC). However, while we can easily distinguish NSCLC from SCLC, it is far less clear to distinct NSCLC subtypes.
- As NSCLC subtypes, there exists two common type which are Lung Adenocarcinoma (**LUAD**) and Lung Squamous Cell Carcinoma (**LUSC**).
  Lung adenocarcinoma starts in glandular cells and is usually located more along the outer edges of the lungs. Lung adenocarcinoma tends to grow more slowly than other lung cancers. Also it accounts for 40% of all lung cancers.
  Lung Squamous Cell Carcinoma begins in the squamous cells which is thin, flat cells that look like fish scales. And it usually occur in the central part of the lung or in one of the main airways. (left or right bronchus)
- To identify candidate progression determinants of NSCLC subtypes, we explored the transcriptomic signatures of **LUAD** versus **LUSC**. We then investigated the prognostic impact of the identified tumor-associated determinants.
- Such disease determinants appeared vastly different in **LUAD** versus **LUSC**, and often had opposite impact on clinical outcome. So we want to find some genes which have an effect on only **LUAD** or **LUSC**.

2. Survival Analysis

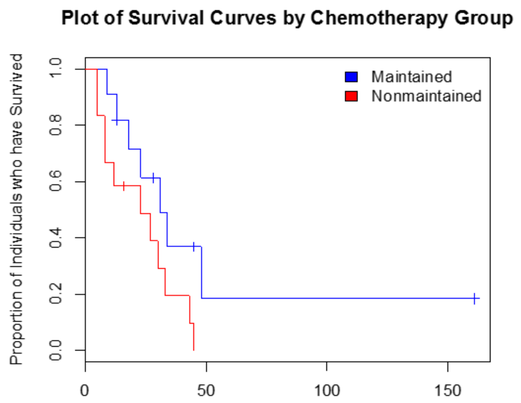
Variable	Description
time	Survival or censoring time
status	censoring status (0 if an individual was censored, 1 otherwise)
x	was maintenance chemotherapy given? ("Maintained" if yes, "Not maintained" if no



Each variables are described in the table, and We can see that the survival times are highly skewed due to the fact that there is a person who survived much longer than everyone lese. In addition, there were quite a few people who wurvived for fewer than 10 years.

In order to determine if there is a statistically significant Between the survival curves, we perform what is known as a **log-rank test**, which tests the following hypothesis

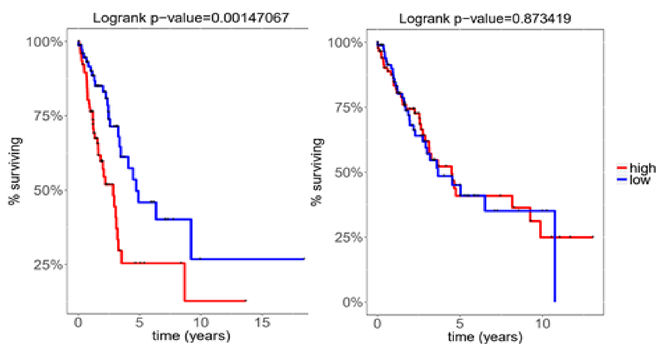
- H_0 : There is no difference in the survival function between those who were on maintenance chemotherapy and those who weren't on maintenance chemotherapy.
- H_a : There is a difference in the survival function between those who were on maintenance chemotherapy and those who weren't on maintenance chemotherapy.



1) Genes which have an effect on **LUAD**

- **ITGA6**

LUAD **LUSC**

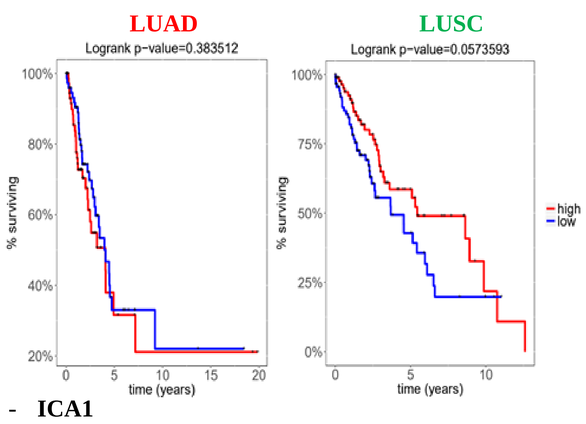


In **LUAD**, Log-rank p-value is so small that we can reject Null hypothesis. However, Log-rank p-value is too big to reject Null hypothesis in **LUSC**.
That is, There is difference between high expression and low expression only in **LUAD**
Also we can see facts just by looking at the picture.

2) Genes which have an effect on **LUSC**

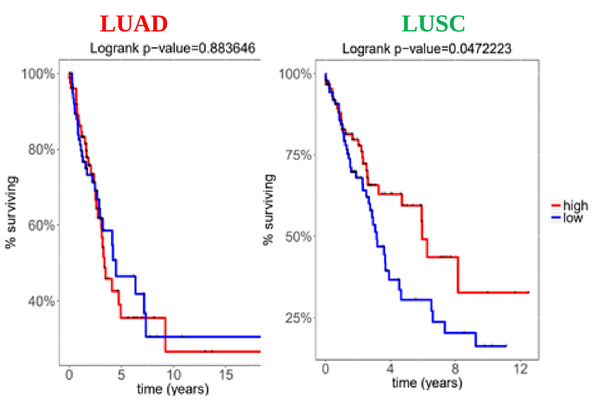
- **ALDOC**

LUAD **LUSC**



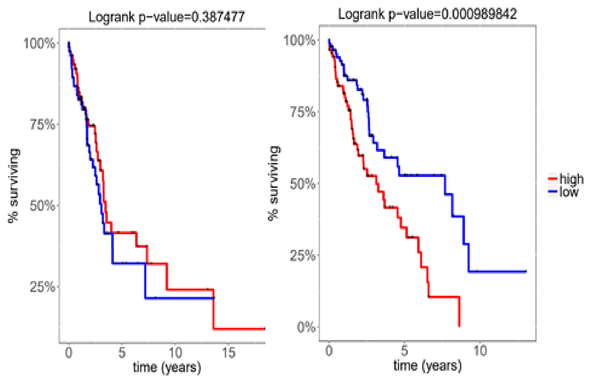
- **CSTA**

LUAD **LUSC**



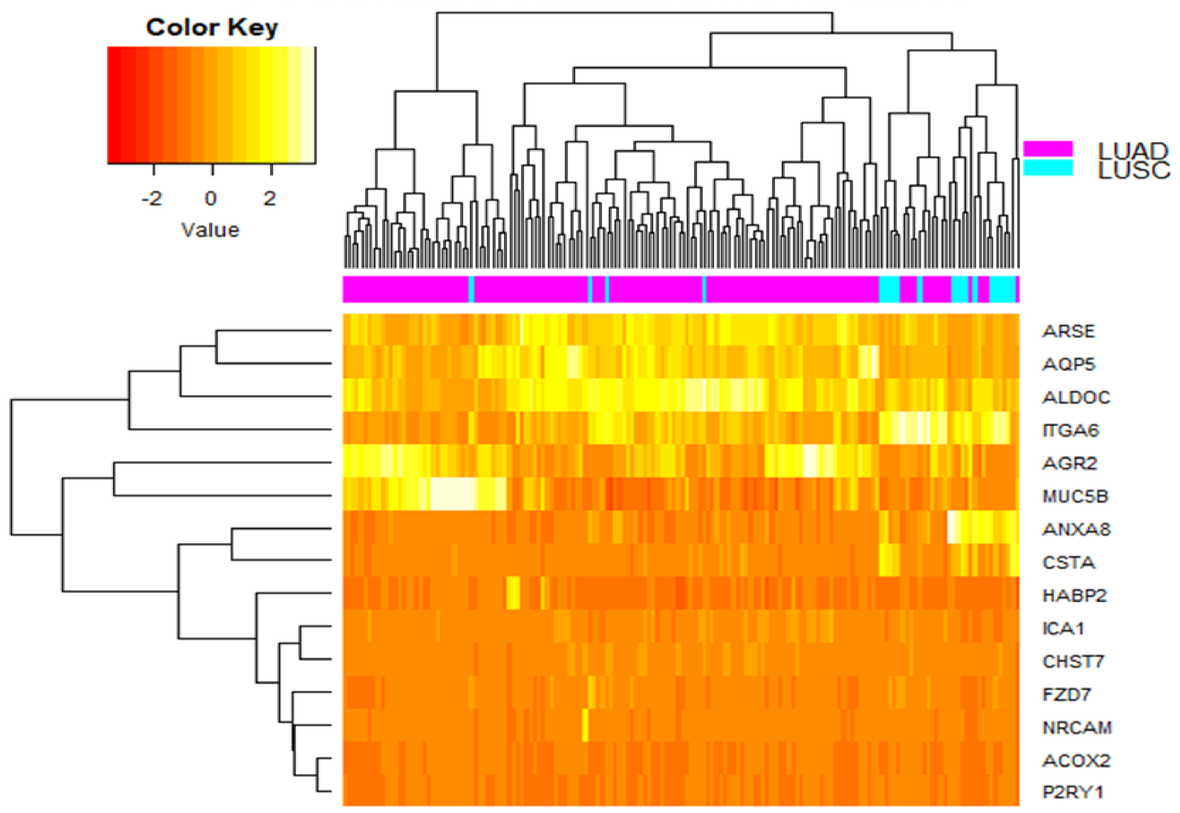
- **ICA1**

LUAD **LUSC**



All of three genes, when we see each survival plot in **LUAD** there is no difference between high expression and low expression. And also p-values are so big that we can not reject Null hypothesis.
In **LUSC**, all of p-values are so small that we can reject Null hypothesis. And all of them have difference between high expression and low expression by looking at the picture. However, while high expression is much larger than low expression in both ALDOC and CSTA, low expression is larger than high expression in ICA1.

3. Clustering Analysis with Heatmap



ITGA6, ANXA8 and CSTA genes are typically divided into **LUAD** and **LUSC**.
In all of three genes, **LUSC** parts are more brighter than **LUAD** parts. But Overall, ITGA6 is brighter than other two genes.
Also we can make two groups. One is made into six genes above and The other is made into next nine genes.

4. Functional Annotations

- Genes which have an effect on **LUAD**

- **ITGA6**

: The gene encodes a member of the integrin alpha chain family of proteins. Integrins are heterodimeric integral membrane proteins composed of an alpha chain and a beta chain that function in cell surface adhesion and signaling. The encoded preproprotein is proteolytically processed to generate light and heavy chains that comprise the alpha 6 subunit. This subunit may associate with a beta 1 or beta 4 subunit to form an integrin that interacts with extracellular matrix proteins including members of the laminin family. The alpha 6 beta 4 integrin may promote tumorigenesis, while the alpha 6 beta 1 integrin may negatively regulate erbB2/HER2 signaling. Alternative splicing results in multiple transcript variants.

- Genes which have an effect on **LUSC**

- **ALDOC**

: This gene encodes a member of the class I fructose-biphosphate aldolase gene family. Expressed specifically in the hippocampus and Purkinje cells of the brain, the encoded protein is a glycolytic enzyme that catalyzes the reversible aldol cleavage of fructose-1,6-biphosphate and fructose 1-phosphate to dihydroxyacetone phosphate and either glyceraldehyde-3-phosphate or glyceraldehyde, respectively.

- **CSTA**

: The cystatin superfamily encompasses proteins that contain multiple cystatin-like sequences. Some of the members are active cysteine protease inhibitors, while others have lost or perhaps never acquired this inhibitory activity. There are three inhibitory families in the superfamily, including the type 1 cystatins (stefins), type 2 cystatins, and kininogens. This gene encodes a stefin that functions as a cysteine protease inhibitor, forming tight complexes with papain and the cathepsins B, H, and L. The protein is one of the precursor proteins of cornified cell envelope in keratinocytes and plays a role in epidermal development and maintenance. Stefins have been proposed as prognostic and diagnostic tools for cancer.

- **ICA1**

: This gene encodes a protein with an arfaptin homology domain that is found both in the cytosol and as membrane-bound form on the Golgi complex and immature secretory granules. This protein is believed to be an autoantigen in insulin-dependent diabetes mellitus and primary Sjogren's syndrome. Several transcript variants encoding two different isoforms have been found for this gene.

5. References

- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6238974/#R2
- https://github.com/ELELAB/LUAD_LUSC_TCGA_comparison
- https://www.genecards.org/cgi-bin/carddisp.pl?gene=ICA1&keywords=ica1
- https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R7_LogisticRegression-Survival/R7_LogisticRegression-Survival4.html