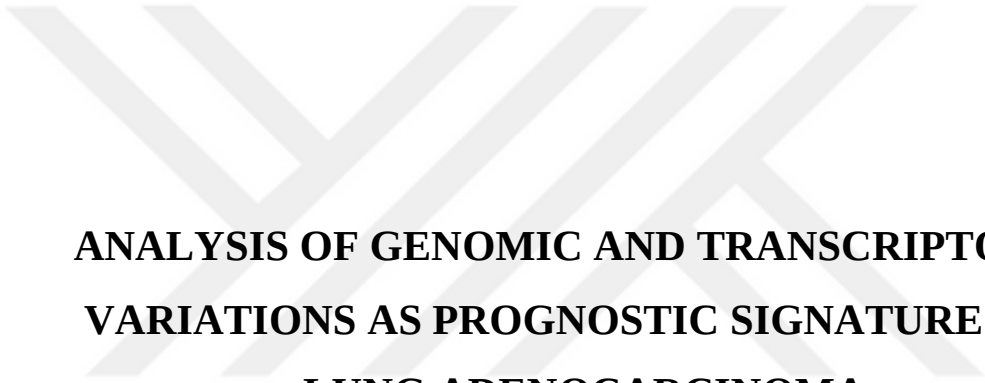


**THE REPUBLIC OF TURKEY
MUĞLA SITKI KOÇMAN UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED
SCIENCES**

DEPARTMENT OF BIOINFORMATICS



**ANALYSIS OF GENOMIC AND TRANSCRIPTOMIC
VARIATIONS AS PROGNOSTIC SIGNATURE FOR
LUNG ADENOCARCINOMA**

MASTER OF SCIENCE

TALİP ZENGİN

**JUNE 2019
MUĞLA**

MUĞLA SITKI KOÇMAN UNIVERSITY
Graduate School of Natural and Applied Sciences

APPROVAL OF THE THESIS

The thesis submitted by **TALİP ZENGİN** with the title of "**ANALYSIS OF GENOMIC AND TRANSCRIPTOMIC VARIATIONS AS PROGNOSTIC SIGNATURE FOR LUNG ADENOCARCINOMA**" has been unanimously accepted by the jury members on the 18th of June 2019 to fulfill the requirements for the degree of Master of Science in the Department of Bioinformatics by the jury listed below.

THESIS JURY MEMBERS

Assist. Prof. Dr. Barış Ethem SÜZEK (**Head of Committee**)

Department of Computer Engineering,
Muğla Sıtkı Koçman University, Muğla

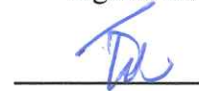
Signature:



Assist. Prof. Dr. Tuğba SÜZEK (**Advisor**)

Department of Computer Engineering,
Muğla Sıtkı Koçman University, Muğla

Signature:



Assist. Prof. Dr. Kürşad TOSUN (**Member**)

Department of Mathematics,
Siena College, New York, USA

Signature:



APPROVAL OF HEAD OF THE DEPARTMENT

Assist. Prof. Dr. Tuğba SÜZEK

Head of Department, Department of Bioinformatics

Signature:



Assist. Prof. Dr. Tuğba SÜZEK

Supervisor, Department of Computer Engineering

Signature:



Date of Defense: 18/06/2019

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.



Talip ZENGİN

18/06/2019

ACKNOWLEDGMENTS

First and the foremost I would like to thank my supervisor Assist. Prof. Dr. Tuğba SÜZEK for her academic guidance and support during my thesis.

I am indebted to Assist. Prof. Dr. Kürşad TOSUN because he taught statistics from bottom-up and introduced R to me. I am also grateful for his nice conversations with coffee.

I am thankful to Assist. Prof. Dr. Barış Ethem SÜZEK for his contributions and comments on my thesis.

I am also very thankful to Jimmy Van den Eynden who is one of the writers of SomInaClust R package, Panos Balomenos who is one of the writers of DEsubs R package, Tiago Chedraoui Silva who is one of the writers of TCGAbiolinks R package and Michael Love who is one of the writers of DESeq2 R package due to their help and suggestions for my analyses.

I thank my ancient and veteran laptop for bearing my two master thesis and heavy computing works for it.

I am also thankful to all people who are supportive of my thesis and Bioinformatics learning.

I am especially thankful to Mehtap ÖZTÜRK for her endless support during my thesis. I also thank Killua and Kömür for not leaving me alone by sleeping on my lap while writing my thesis.

My deepest gratitude goes to my family for their unflagging love, support and prayers throughout my life.

The numerical calculations reported in this thesis were fully performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources).

ÖZET

AKCİĞER ADENOKARSİNOM İÇİN PROGNOSTİK İMZA OLARAK KULLANILABİLECEK GENOMİK VE TRANSKRİPTOMİK VARYASYONLARIN ANALİZİ

Talip ZENGİN

Yüksek Lisans Tezi

Fen Bilimleri Enstitüsü

Biyoinformatik Anabilim Dalı

Danışman: Dr. Öğretim Üyesi Tuğba SÜZEK

Haziran 2019, 134 sayfa

Akciğer kanseri, dünya çapında en fazla sayıda kanser kaynaklı ölüme sebep olmaktadır ve akciğer adenokarsinomu (LUAD) akciğer kanserinin en yaygın görülen tipidir. Bu çalışmada, LUAD hastalarının, Kanser Genom Atlas'ından (TCGA) indirilen Tek-nükleotid varyasyonları (SNV'ler), kopya sayısı varyasyonları (CNV'ler), RNA sekansı ve klinik verilerini içeren entegre bir meta-analiz gerçekleştirildi. Analizde tespit edilen anlamlı varyasyonların Cox Oransal Tehlikeler Modeli aracılığıyla genel sağ kalım üzerindeki etkileri araştırıldı. 55 LUAD hastasının aktif alt ağlarında, 25 farklı ifade edilmiş genin (DEG) ifadesi ile önemli ölçüde mutasyona uğramış 4 gen saptandı. Ayrıca 303 CNV, 41 SNV, 8 aktif alt ağ DEG ve tümör evresi, 510 LUAD hastasının genel sağ kalımına etki eden değişkenler olarak tanımlandı. Bu tanımlanmış genlerin yolaklarını analiz ettiğimizde, yolaklarının yüksek oranda kanser ve akciğer kanseri ile ilişkili olduğu bulundu. Bu etkili genlerden bazıları da bağışıklık sistemi yolaklarında rol almaktadır. Bu DEG'lerin ısı haritası belirgin ikiye bölünmüş bir desen göstermiş ve hastalar hiyerarşik kümeleme yoluyla iki gruba kümelendiklerinde ve sağ kalım analizleri yapıldığında, sağ kalım grafiği, iki hasta kümesinin sağ kalım olasılığı arasında anlamlı bir farklılık göstermiştir. Tespit edilen genlerin birçoğu iyi bilinen akciğer kanseri ile ilişkili genlerdir ve bazıları yakın zamanda akciğer kanseri ile ilişkilendirilmiştir. Ancak, PRKCE, RAPGEF3, COL6A5, ROBO2, CACNA1D gibi bazı genler ile akciğer kanseri henüz ilişkilendirilmemiştir. Tanımlanan tüm bu genler, LUAD için klinik öngörü için moleküler imza olarak kullanılabilecek güçlü adaylardır.

Anahtar Kelimeler: TCGA, LUAD, SNV, CNV, RNAsekans, Aktif alt ağ, Cox Oransal Tehlikeler Modeli, Hiyerarşik Kümeleme, Isı Haritası, Sağ Kalım Analizi

ABSTRACT

ANALYSIS OF GENOMIC AND TRANSCRIPTOMIC VARIATIONS AS PROGNOSTIC SIGNATURE FOR LUNG ADENOCARCINOMA

Talip ZENGİN

Master of Science (M.Sc.)

Graduate School of Natural and Applied Sciences

Department of Bioinformatics

Supervisor: Assist. Prof. Dr. Tuğba SÜZEK

June 2019, 134 pages

Lung cancer is the leading cause of the largest number of deaths worldwide and lung adenocarcinoma (LUAD) is the most common form of lung cancer. In this study, we carried out an integrated meta-analysis of the mutations including single-nucleotide variations (SNVs), the copy number variations (CNVs), RNA-seq and clinical data of LUAD patients downloaded from The Cancer Genome Atlas (TCGA) and reported their influence on overall survival through Cox Proportional Hazards Model. We identified 4 significantly mutated genes with the expression of 25 differentially expressed genes (DEGs) at active subnetworks of 55 LUAD patients have a high impact on the survival of the patients. We also identified 303 CNVs, 41 SNVs, 8 active subnetwork DEGs and tumor stage as impactful variates on overall survival of 510 LUAD patients. When we analyze the pathways of these identified genes, it is found that their pathways are highly cancer and lung cancer-related. There are also dysregulated immune system pathways of these impactful genes. The heatmap of these DEGs showed an apparent dichotomous pattern and when the patients were clustered into two by hierarchical clustering and survival analysis was performed, survival plot showed significant divergence between survival probability of two clusters of patients. Many of the identified impactful genes are well-known lung cancer-related genes, and some of them are recently identified lung cancer-related genes. But there is no evidence of direct interaction between impactful genes such as PRKCE, RAPGEF3, COL6A5, ROBO2, CACNA1D genes, and lung cancer, yet. These identified impactful genes are strong candidates as a molecular signature for clinical prediction for LUAD.

Keywords: TCGA, LUAD, SNV, CNV, RNAseq, Active Subnetwork, Cox Proportional Hazards Model, Hierarchical Clustering, Heatmap, Survival Analysis



To my father

TABLE OF CONTENTS

ACKNOWLEDGEMENTS.....	iv
ÖZET.....	v
ABSTRACT.....	vi
TABLE OF CONTENTS.....	viii
LIST OF TABLES	x
LIST OF FIGURES	xi
1. INTRODUCTION.....	xi
1.1. Cancer as a Complex Disease	1
1.2. Cancer Genome Analysis.....	2
1.3. Integrative Cancer Genome Analysis	4
1.4. Clinical Impacts of Cancer Genome Analysis	6
1.5. Lung Cancer	7
1.6. The Cancer Genome Atlas (TCGA).....	10
2. METHODS	12
2.1. Data	12
2.2. Identification of Significant Simple Nucleotide Variations.....	12
2.3. Identification of Significant Copy Number Variations.....	13
2.4. Differential Expression Analysis (DEA)	14
2.5. Active Subnetwork and Pathway Analysis	14
2.6. Statistical Analysis.....	14
3. RESULTS	16
3.1. Identification of Significant Simple Nucleotide Variations.....	16
3.1.1. SNV analysis of 55 LUAD patients.....	16
3.1.1.1. Determination of important mutated genes by SomInaClust package	19
3.1.2. SNV analysis of 510 LUAD patients.....	21
3.1.2.1. Determination of important mutated genes by SomInaClust package	23
3.2. Identification of Significant Copy Number Variations.....	27
3.2.1. CNV analysis of 55 LUAD patients	27
3.2.2. CNV analysis of 510 LUAD patients	31

3.3. Differential Expression Analysis (DEA)	35
3.3.1. DEA of 55 LUAD patients.....	35
3.3.2. DEA of 510 LUAD patients.....	40
3.4. Active Subnetwork and Pathway Analysis	45
3.4.1. Active subnetwork and pathway analysis of 55 LUAD patients	45
3.4.2. Active subnetwork and pathway analysis of 510 LUAD patients	51
3.5. Statistical Analysis.....	56
3.5.1. Statistical analysis of 55 LUAD patients	57
3.5.2. Statistical analysis of 510 LUAD patients	62
4. DISCUSSION	76
5. CONCLUSION.....	84
REFERENCES.....	85
APPENDIX A.	92
CIRCULUM VITAE.....	133

LIST OF TABLES

Table 3.1. SomInaClust results showing significantly mutated genes in 55 LUAD patients.....	20
Table 3.2. SomInaClust results showing significantly mutated genes in 510 LUAD patients.....	25
Table 3.3. Top ten significant amplified and deleted genes in 55 LUAD patients....	28
Table 3.4. Top ten significant amplified and deleted genes in 510 LUAD patients...	32
Table 3.5. Top ten significant down-regulated genes in tumor samples of 55 LUAD patients.....	36
Table 3.6. Top ten significant up-regulated genes in tumor samples of 55 LUAD patients.....	37
Table 3.7. Top ten significant down-regulated genes in tumor samples of 510 LUAD patients.....	41
Table 3.8. Top ten significant up-regulated genes in tumor samples of 510 LUAD patients.....	42
Table 3.9. Summary of clinical features of 55 LUAD patients.	57
Table 3.10. Candidate driver genes determined by multivariate CPHR, of 55 LUAD patients	59
Table 3.11. Summary of clinical features of 510 LUAD patients	62
Table 3.12. Impactful CNVs determined by multivariate CPHR, of 510 LUAD patients	64
Table 3.13. Impactful SNVs determined by multivariate CPHR, of 510 LUAD patients	72
Table 3.14. Impactful DEGs and features determined by mv CPHR, of 510 LUAD patients.	73

LIST OF FIGURES

Figure 2.1. The flowchart of the pipeline.....	15
Figure 3.1. Summary plot for mutation data of 55 LUAD patients	17
Figure 3.2. Oncoplot showing top ten mutated genes of tumor samples of 55 LUAD patients	18
Figure 3.3. Transition and transversion rates in tumor samples of 55 LUAD patients.	19
Figure 3.4. Pyramid plot of SomInaClust results in tumor samples of 55 LUAD patients.	20
Figure 3.5. Summary plot for mutation data of 510 patients in LUAD project.....	21
Figure 3.6. Oncoplot showing top ten mutated genes of tumor samples of 510 LUAD patients.	22
Figure 3.7. Transition and transversion rates in tumor samples of 510 LUAD patients.	23
Figure 3.8. Pyramid plot of SomInaClust results in tumor samples of 510 LUAD patients	24
Figure 3.9. Recurrent CNVs on all chromosomes in tumor samples of 55 LUAD patients.	28
Figure 3.10. Pathways of significant amplification in tumor samples of 55 LUAD patients.....	30
Figure 3.11. Pathways of significant deletions in tumor samples of 55 LUAD patients	30
Figure 3.12. Recurrent CNVs on all chromosomes in tumor samples of 510 LUAD patients.....	31
Figure 3.13. Top 20 Pathways of significant aberrations in tumor samples of 510 LUAD patients.....	33
Figure 3.14. Top 20 pathways of significant amplification in tumor samples of 510 LUAD	34

Figure 3.15. Top 20 Pathways of significant deletions in tumor samples of 510 LUAD patients.	34
Figure 3.16. Volcano plot of differentially expressed genes in tumor samples of 55 LUAD patients.	36
Figure 3.17. Top 20 dysregulated pathways of gene expression in tumor samples of 55 LUAD patients.	38
Figure 3.18. Upregulated pathways of gene expression in tumor samples of 55 LUAD patients.	39
Figure 3.19. Downregulated pathways of gene expression in tumor samples of 55 LUAD patients.	40
Figure 3.20. Volcano plot of differentially expressed genes in tumor samples of 510 LUAD patients.	41
Figure 3.21. Top 20 dysregulated pathways of gene expression in 510 LUAD patients.	43
Figure 3.22. Upregulated pathways of gene expression in tumor samples of 510 LUAD patients.	44
Figure 3.23. Downregulated pathways of gene expression in tumor samples of 510 LUAD patients.	45
Figure 3.24. Top 10 significant subnetwork genes in tumor samples of 55 LUAD patients.	46
Figure 3.25. Heat map showing functional measurements of top 10 subnetwork genes in tumor samples of 55 LUAD patients.	46
Figure 3.26. Graph of subnetwork 1 from 55 LUAD patients.	47
Figure 3.27. Graph of subnetwork 2 from 55 LUAD patients.	47
Figure 3.28. Graph of subnetwork 3 from 55 LUAD patients.	48
Figure 3.29. KEGG pathways which include subnetwork genes of 55 LUAD patients.	49
Figure 3.30. TFs which are involved or related to subnetworks of 55 LUAD patients.	50
Figure 3.31. Drugs which are related to proteins in subnetworks of 55 LUAD patients.	50

Figure 3.32. Top 10 significant subnetwork genes in tumor samples of 510 LUAD patients.	51
Figure 3.33. Heat map showing functional measurements of top 10 subnetwork genes in tumor samples of 510 LUAD patients	52
Figure 3.34. Graph of subnetwork 1 from 510 LUAD patients.	52
Figure 3.35. Graph of subnetwork 2 from 510 LUAD patients	53
Figure 3.36. Graph of subnetwork 3 from 510 LUAD patients.	53
Figure 3.37. KEGG pathways which include subnetwork genes of 510 LUAD patients.	54
Figure 3.38. TFs which are involved or related to subnetworks of 510 LUAD patients	55
Figure 3.39. Drugs which are related to proteins in subnetworks of 510 LUAD patients.	56
Figure 3.40. Overall survival plot for tumor stages of tumor samples of 55 LUAD patients	58
Figure 3.41. Pathways of significant genes determined by multivariate CPHR, of 55 LUAD patients.	60
Figure 3.42. Heatmap of significant DEGs from multivariate CPHR for 55 LUAD patients.	61
Figure 3.43. Overall survival plot of hierarchical clusters of 55 LUAD patients.....	61
Figure 3.44. Overall survival plot for tumor stages of tumor samples of 510 LUAD patients.	63
Figure 3.45. Pathways of significant genes determined by multivariate CPHR, of 510 LUAD patients	74
Figure 3.46. Heatmap of significant DEGs from multivariate CPHR for 510 LUAD patients.	75
Figure 3.47. Overall survival plot of hierarchical clusters of 510 LUAD patients....	75

1. INTRODUCTION

1.1. Cancer as a Complex Disease

Genetic diseases are associated with some sort of mutations in DNA or RNA or protein that affects normal functionality. In order to find disease-causing mutations, polymorphisms in genome of patients which are absent in healthy people should be detected (Bromberg, 2013). Cancer is a complex disease in which many mutated genes play a role instead of a single gene mutation. Therefore, cancer is defined as a “disease of genes” and this nature of cancer makes detecting and characterizing the cancer-related mutations important for cancer therapy. The recent developments in sequencing methods allow large scale genome projects that systematically analyze the genomic and transcriptomic profile of cancer cells (Vazquez, 2012).

Cancer is the final stage of abnormal cells which gain accumulation of somatic mutations which are genomic variants such as point mutations, deletions, insertions and copy number variations of non-germ line cells at the tumor site or the affected tissues, during the lifetime (Vazquez, 2012). Single nucleotide polymorphisms (SNPs) are point mutations (single base-pair changes) and can have different functional consequences such as changing of binding affinity of transcription factor proteins on DNA regions, mRNA stability or amino acid sequences of proteins. The mostly observed genomic variants are SNPs in the human genome and they occur with high frequency (Bush & Moore, 2012). An individual may contain about 10-15 million SNPs in the genome. However, not all SNPs are associated with harmful effects (Bromberg, 2013). SNPs may cause silent mutation (no amino acid change on protein sequence) due to the variety of aminoacid codes, and neutral mutation (no effect on protein function), or missense mutation (single amino acid substitution), frame-shift

mutation (results in completely different protein sequence), nonsense mutation (results in premature stop codon). Most of the observed mutations do not show phenotypic differences, some are weakly harmful according to normal function and some of them are still weakly beneficial (Bromberg, 2013). Copy number variations (CNVs) have also important roles in cancer by activating oncogenes and/or by inactivating tumor suppressors. Therefore, characterizing the effects of somatic CNVs is important for cancer diagnostics and therapy (Zack, 2013). A small portion of mutations is strongly deleterious or inactivating mutations which affect protein function. But complex disorders can accumulate weakly deleterious mutations that cause abnormal phenotype. The cancer genome consists of many mutations, some of them are fundamental for tumorigenesis, called as ‘driver mutations’, and others occur during the growth of the tumor, called as ‘passenger mutations’ (Vazquez, 2012). Therefore, it is very important to determine driver and passenger mutations of tumorigenesis and tumor growth for cancer therapy (Bromberg, 2013).

1.2. Cancer Genome Analysis

There are different steps in cancer genome analysis in which different research areas are involved, different bioinformatics methods are performed and different challenges are existing. In the beginning, the genome of tumor samples is sequenced through newly developed techniques such as Next Generation Sequencing (NGS) and then reads from small pieces of DNA are collocated by aligning to the sequence of the reference genome. Then genomic variants and their genomic locations are determined by the variant calling process. The workflow for this first level analysis is called primary analysis (Vazquez, 2012). The genomic differences obtained from the primary analysis are carefully analyzed to determine the effects of genomic variations. Mutations in DNA can potentially alter the amino acid sequence of proteins. The impact of the amino acid variations may range from mostly irrelevant to highly deleterious on the function of the protein. The severity of the alterations can be

determined by using specialized software tools. Genomic variations are also analyzed to determine the frequency of mutations, which can help to identify crucial genes and hot spots of mutations. In order to identify potential ‘driver mutations’ which play a role directly in tumorigenesis, the predicted outcome and the frequency of the specific mutations are also analyzed (Vazquez, 2012). It is easy to identify the recurrently mutated genes in cancer such as KRAS and TP53 which are examples of mostly mutated genes in many cancer types. On the other hand, the possibility of detecting the same mutation in different cancer samples is low because of the distribution of the mutations. It is more possible to determine common functional alterations associated with cancer of interest. Therefore, pathway analysis can be very advantageous by determining the cell signaling pathways, protein complexes, functional classes of proteins, regulatory protein networks or clusters in protein-protein interaction networks, associated with mutated genes. Pathway analysis has become one of the fundamental components of cancer genome analysis. Therefore, cancer can also be defined as ‘disease of the pathways’ through the mutated genes which play a role in signaling pathways (Vazquez, 2012).

Biological interaction networks resulted from complex interactions between proteins, genes, RNAs, epigenetic factors, metabolites, cell signaling stimulators, cells, and tissues, predestine of the health (Kristensen, 2014). Cancer is the result of genomic, transcriptomic or proteomic alterations in a complex network of signaling pathways causing abnormal cell behaviors. The complexity of protein interaction networks is caused by a high number of components, multiple connections between pathways and many feedback loops (Laubenbacher, 2009). Moreover, the proteins of each signaling pathway play a role at particular levels which changes continuously and dynamically in response to stimulators and differential cell-cell interactions. Cell function and behaviors are affected by this complex and dynamic signaling network through active sub-networks that control different functions or behaviors of the cell. It is needed to understand the signaling networks involved in carcinogenesis in order to advance the knowledge of tumorigenesis, tumor growth and metastasis. Alterations in signaling

networks that are resulted from genetic, epigenetic and environmental changes, accumulate at each stage of tumorigenesis (Thorgeirsson, 2002). And, the alterations in specific signaling networks generate the hallmarks of cancer which are mimicking normal cell growth signaling, the high-level potential of replication, blocking anti-growth signals, bypassing apoptosis, constant angiogenesis, invasion and distant metastasis (Hanahan & Weinberg, 2011; Laubenbacher, 2009). Signaling network analysis is also beneficial in clinical diagnosis, detection of key proteins of cancer, development of drugs and cancer therapies (Zhang, 2009).

1.3. Integrative Cancer Genome Analysis

Genome analysis alone is not enough and other information about protein interactions contributing to tumor initiation and development are required to develop cancer models (Hanash, 2004). Integration, visualization and disease-centered interpretation of different data are required for detailed cancer genome analysis (Vazquez, 2012). Most tumors consist of large numbers of genomic alterations, but it is hard to determine the driver genes out of them. New strategies are available for global profiling of changes at genomic, transcriptomic and proteomic levels basically. The genome of patients can be analyzed for genomic variations, such as point mutations, deletions, insertions, copy number variations (CNVs) or differential DNA methylation by NGS. And also, the tumor transcriptome can be analyzed by DNA microarrays, SAGE or RNAseq techniques for gene expression profiling. And also, tumor proteome can be analyzed by protein microarrays, Immuno-histo/cyto-chemistry techniques, or mass spectrometry to detect expression levels and alterations in proteins that may not be identified by genome and transcriptome analysis (Hanash, 2004). Some of the genome alterations such as copy number variations (CNVs), or mutations that affect the binding affinity of transcription factors, or mutations in promoter, silencer and enhancer regions, or differential status of epigenetics, can affect the gene expression profile. Therefore, genome analysis should be integrated

with transcriptome or proteome profiling to understand the biology of cancer at least two main levels (Hanash, 2004). This type of study requires to combine more than a few combinations of such data and it needs special applications to work on specific data sets (Vazquez, 2012).

Recent cancer studies have integrated different types of data sets generated by different global profiling techniques. Mostly two types of data sets (genomics and epigenomics; genomics and transcriptomics; transcriptomics and proteomics) from the same patient samples are combined for analysis. Clinical and pathological characteristics of patients have been also considered for these integrated data sets (Hanash, 2004). As an example, Hicks et al. presented an integrative genomics approach combining genomic and transcriptomic data to study the effect of combinations of multiple genomic alterations in genes and signaling pathways. They showed that the genes which are related to breast cancer risk and containing SNPs play a role in interaction at signaling pathways associated with breast cancer and they are functionally related (Hicks, 2011). At the next research Hicks et al. determined mutated genes and signaling pathways in prostate cancer such as JAK2, IGF-1, prolactin and androgen signaling pathways (Hicks, 2014). Zhang et al. designed a method for integrating genomic and epigenomics data to identify key characteristics associated with ovarian cancer (Zhang, 2014). This method aims to determine the relationships between key characteristics through probabilistic graphical mode by using The Cancer Genome Atlas (TCGA) data. They showed that there are 13 hub genes including CSKN2A1, ARID1A, COL5A2, and C19orf53. They found four gene clusters related to cell proliferation, invasion, and mitochondrial pathways. And two genes of glycoprotein synthesis system, PSG11, and GALNT10, were significantly predictive for the overall survival time. Zhang et al. integrated somatic mutation, transcriptome profile and clinical data of breast cancer patients to identify functional gene clusters which can be used for grouping of patients based on clinical outcomes. They used transcriptome profile in order to group the patients by using functionally related mutation targeted proteins. They determined mutated genes which can be used

for potential biomarkers related to patient survival and a stable subgroup of patients who have ER (-) and triple-negative breast cancer subtypes (Zhang, 2016). All results above show that it is important to perform integrative data analysis to enlighten the molecular nature of the cancer cells.

1.4. Clinical Impacts of Cancer Genome Analysis

Cancer genome analysis can be performed on both the cancer type and the patient. Research on cancer types is used to identify cancer biomarkers, to understand tumorigenesis process or to characterize cancer subgroups showing the different clinical outcome, by using the integrated biological and clinical data of a group of patients with a specific cancer type. Genome analysis of a cancer patient is performed to identify genomic variations which may be used for specifically designed cancer treatment (Vazquez, 2012). The information generated from cancer genome analysis performed on the cancer type can be used for the determination of cancer type and analysis of biomarkers of tumor of patients. Therefore, the best therapeutics can be determined specifically for a patient. In regular clinic applications, patients are grouped based on characteristics of the patient (age, smoking status, and performance status) and the tumor (tumor size, histological stage, the expression pattern of a few proteins, and metastasis status). However, this type of grouping does not consider the molecular characteristics of the tumor. A substantial number of patients are over-treated, present drug resistance chemotherapy or get exposed to the toxic side effects of the therapeutics without deriving any significant benefit. Gene-expression profiling and several genomic tests have been developed to determine the therapeutic and to anticipate clinical outcome (Cardoso, 2016). One such test, the 70-gene signature (MammaPrint), classifies tumors into two groups good prognosis and poor prognosis depending on the previous data about the risk of recurrence at 5 years and at 10 years (van 't Veer, 2002). Cardoso, F. et al. detected the genomic risk by MammaPrint method and clinical risk of 6693 women with early-stage breast cancer, in a

randomized, phase III study. The women at low genomic and clinical risk did not take chemotherapy, while women at high genomic and clinical risk took chemotherapy. MammaPrint study showed that 5-year survival rate without distant metastasis for patients who have a high clinical risk with low genomic risk and received chemotherapy was 1.5% insignificantly higher than the patients who were in the same group and did not receive chemotherapy. These results show that about 46% of patients who have a high clinical risk with low genomic risk may not need chemotherapy at all (Cardoso, 2016). This clinical study supports the importance of gene expression and biomarker information on characterizing the molecular nature of the tumor and determining the best therapy type for that patient specifically. But the gene expression is not enough alone and we need to combine different -omics data to analyze the biological characteristics of the tumor of a patient. Only with the integration of the patient's full molecular and clinical profile, better therapy types can be tailored to cure each individual cancer patient.

1.5. Lung Cancer

Lung cancer is the most common cancer and responsible for the largest number of deaths worldwide with 1.8 million deaths, 18.4% of the total (IARC, 2018). Lung cancer is categorized into two main categories: non-small cell lung cancer (NSCLC) which occurs in 85% of patients and small cell lung cancer (SCLC) in 15% of cases. NSCLC is grouped into 3 histological sub-types: lung adenocarcinoma (LUAD) which is the most common form of lung cancer, lung squamous cell carcinoma (LUSC) and large cell carcinoma (Travis, 2011). Multiple genomics and transcriptomics researches have been performed to understand the molecular biology of LUAD tumor samples. Whole Exome Sequencing (WES) and targeted sequencing of protein-coding regions of tumor and normal samples have determined more than 20 mutated genes related with LUAD (Ding, 2008; Govindan., 2012; Imielinski, 2012; Kan, 2010; Liu, 2012). RNA sequencing has identified gene fusions, such as KIF5B-

RET (Ju, 2012; Kohno, 2012), EML4-ALK (Aburatani, 2007) and ROS1 fusions (Togashi, 2012). Array-based copy number variation studies (The Cancer Genome Atlas Research Network, 2014; Weir, 2007) have discovered that some genes are highly amplified genes, such as EGFR, NKX2-1, MET, and TERT and some genes are deleted such as CDKN2A. These studies determined some important biomarkers which can be molecular targets for cancer therapy. Recently developed targeted kinase inhibitors are more efficient than traditional chemotherapy drugs for EGFR mutations or ALK fusions (Lindeman, 2013). Although many driver genes have been identified to target for cancer therapy, there is still a need to analyze more LUAD tumor samples to identify additional key genes and signaling pathways (Shi, 2016).

Integration of different types of molecular data has been used to characterize the molecular basis of lung cancer and to determine the clinical status of patients. For example, Aramburu et al. performed an integrative analysis by using copy number variations (CNVs) and gene expression profile with clinical characteristics, to identify prognostic genomic status. They generated two candidate gene list for squamous cell carcinoma (SCC) and lung adenocarcinoma (LUAD) including 150 genes and 513 genes respectively, which showed a significant relationship between gene expression and gene copy number, together with prognostic values. By combining these genes and clinical data, hazard ratios were calculated. And they presented a methodology using tumor genomic data and clinical data to anticipate survival time of early staged LUAD and SCC patients (Aramburu, 2015). At another study Huang et al. analyzed the 129 Lung Squamous Cell Carcinoma samples from TCGA by using DNA methylation, gene expression, miRNA expression, and copy number variation (CNV) data. They built the lung cancer protein interaction network by using the known genes related to lung cancer. Through this interaction network, the regulations between genes, miRNAs, DNA methylations, and CNVs were determined. They identified and analyzed novel key role players such as miRNA-31 and miRNA-let-7a; the methylation of HOXD3 and ARHGDIB genes; copy number variation of AGAP2, showing significant relation with the survival time of patients with lung cancer

(Huang, 2015). Shi et al. analyzed 101 LUAD samples by using data from different levels -DNA mutations, gene expression profile, copy number variations, and DNA methylation- in order to identify the relation between genomic status and clinical status. They determined deleterious mutations at ZKSCAN1 and POU4F2 genes which are two novel candidate driver genes (Shi, 2016). Furthermore, recent studies have performed to generate new methods to analyze integrative cancer data. Berger et al. presented a method called expression-based variant-impact phenotyping (eVIP). This method uses gene expression data to determine driver mutations which are impactful and eliminate passenger mutations which are neutral for the cell. They characterized 194 somatic mutations related to primary LUAD. And eVIP determined that 69% of mutations were impactful while 31% of mutations were functionally neutral. They found important impactful mutations such as EGFR (p.S645C), ERBB2 (p.S418T), ARAF (p.S214C) and ARAF (p.S214F) which are examples of rare somatic mutations (Berger, 2016). TCGA network itself analyzed 230 LUAD samples using mRNA, miRNA and DNA sequencing integrated with copy number, methylation and proteomic data (The Cancer Genome Atlas Research Network, 2014) and reported the samples with high rates of somatic mutation. Eighteen genes with high mutation load were reported such as RIT1 activating mutations and MGA loss-of-function mutations. They also identified aberrations in NF1, MET, ERBB2 and RIT1 occurred in 13% of cases and MAPK and PI3K pathway activity (The Cancer Genome Atlas Research Network, 2014). Deng Z. et al. presented genomic alterations in LUAD samples from TCGA and found the significantly aberrant CNV segments which are associated with the immune system and 63 mutated genes associated with lung cancer signaling related to cancer progression. They identified the mutations of the PI3K protein family members include PIK3C2B, PIK3CA, PIK3R1 (Deng, 2017).

1.6. The Cancer Genome Atlas (TCGA)

The Cancer Genome Atlas (TCGA) database (Chang, 2013) is one of the largest and most commonly used public resource assembled through The TCGA Pan-Cancer project, providing data-sets of single nucleotide variation, gene expression quantification, miRNA expression quantification, DNA methylation, copy number variation and reverse-phase protein array with clinical and biospecimen supplement data from thousands of cancer patients. TCGA includes data of 39 tumor types: Neuroblastoma (NBL), Breast Invasive Carcinoma (BRCA), Acute Myeloid Leukemia (AML), High-Risk Wilms Tumor (WT), Glioblastoma Multiforme (GBM), Ovarian Serous Cystadenocarcinoma (OV), Lung Adenocarcinoma (LUAD), Uterine Corpus Endometrial Carcinoma (UCEC), Kidney Renal Clear Cell Carcinoma (KIRC), Head and Neck Squamous Cell Carcinoma (HNSC), Brain Lower Grade Glioma (LGG), Thyroid Carcinoma (THCA), Lung Squamous Cell Carcinoma (LUSC), Prostate Adenocarcinoma (PRAD), Skin Cutaneous Melanoma (SKCM), Colon Adenocarcinoma (COAD), Stomach Adenocarcinoma (STAD), Bladder Urothelial Carcinoma (BLCA), Osteosarcoma (OS), Liver Hepatocellular Carcinoma (LIHC), Cervical Squamous Cell Carcinoma (CESC), Kidney Renal Papillary Cell Carcinoma (KIRP), Sarcoma (SARC), Acute Myeloid Leukemia (LAML), Pancreatic Adenocarcinoma (PAAD), Esophageal Carcinoma (ESCA), Pheochromocytoma and Paraganglioma (PCPG), Rectum Adenocarcinoma (READ), Testicular Germ Cell Tumors (TGCT), Thymoma (THYM), Kidney Chromophobe (KICH), Adrenocortical Carcinoma (ACC), Mesothelioma (MESO), Uveal Melanoma (UVM), Rhabdoid Tumor (RT), Lymphoid Neoplasm Diffuse Large B Cell Lymphoma (DLBC), Uterine Carcinosarcoma (UCS), Cholangiocarcinoma (CHOL), Clear Cell Sarcoma of the Kidney (CCSK) which are listed from higher cases to lower cases respectively in the data portal of TCGA.

We analyzed single-nucleotide variations (SNVs), the copy number variations (CNVs) and RNA-seq profiles of LUAD patients from TCGA and their influence on

overall survival through the Cox Proportional Hazards Model. We identified significantly mutated genes and copy number alterations, with significant differentially expressed genes playing a role in active subnetworks, and their impact on the overall survival of patients. These genes are strong candidates to be used as a molecular signature for clinical prediction or personalized treatment.



2. METHODS

2.1. Data

Simple Nucleotide Variations (SNVs), Transcriptome Profiling, Copy Number Variations (CNVs), and Clinical data of 510 patients (Table 1) who have all these types of data, were downloaded from TCGA harmonized database by using the R/Bioconductor *TCGABiolinks* package (Colaprico, 2016). Only 55 patients (Table 2) of 510 LUAD patients have paired (tumor and matched normal sample) RNA-seq data. We performed two analysis for 55 patients (paired) and 510 patients (unpaired) separately by using their genomic alteration data including Simple Nucleotide Variations, Copy Number Variations, Transcriptome Profiling, processed using the reference of hg38, from the TCGA harmonized archive.

Some patients may have more than one sample. Therefore, in order to unify and choose the best samples of SNVs, CNVs and Transcriptome Profiling data, we used an R code (*tcga_replicateFilter.R*) which was written by Shixiang Wang and presented at *GitHub*. This code filters the replicates in sample barcodes of patients who have more than one sample in analysis. You can find this code with all other codes used for this project at Appendix.

2.2. Identification of Significant Simple Nucleotide Variations

The Mutation Annotation Format (maf) file contained somatic mutations of all patients in the TCGA LUAD project, was downloaded using *TCGABiolinks* package. The other R/Bioconductor package, *maftools* (Mayakonda, 2018), were used to subset original maf file by tumor sample barcodes of patients of interest. *Maftools* package

also summarizes the mutations and represents a summary plot, oncoplot, and transition and transversion plot. Significant mutations among tumor samples were divided into two groups, oncogene (OG) or tumor suppressor gene (TSG) by using R package *SomInaClust* (Van den Eynden, 2015). *SomInaClust* works on the basic assumption that driver genes in tumor samples have clustered mutations or a high number of inactivating mutations because of the selective pressure during tumorigenesis. Based on this assumption, while oncogenes have clustered mutations, tumor suppressors have inactivating (protein-truncating) mutations such as nonsense mutations and frameshift insertions/deletions. *SomInaClust* uses a reference step in which background mutation rate and hot spots are determined for genes existing in reference mutation database such as *COSMIC* (v88 was used for this study) database (Forbes, 2017).

2.3. Identification of Significant Copy Number Variations

The CNV dataset for primary solid tumor samples, generated by Affymetrix Genome-Wide Human SNP Array 6.0 platform, was downloaded using *TCGAbiolinks* package. The significant aberrant genomic regions were identified by R/Bioconductor *GAIA* package (Morganella, 2011). *GAIA* (Genomic Analysis of Important Alterations) package has procedure based on two main steps called significance testing and homogeneous peel-off. At the first step, it determines the significant genomic aberrations among samples. The peel-off step identifies the significant variation values corrected by FDR for a region above the significance threshold (Morganella, 2011). NCBI IDs and Hugo Symbols of the genes with differential copy number were determined using *biomaRt* package (Durinck, 2009).

2.4. Differential Expression Analysis (DEA)

The Transcriptome Profiling data of LUAD patients in mRNA expression level (as unnormalized HTSeq raw counts), generated by TCGA RNA-seq pipeline was downloaded by *TCGABiolinks* package. Differentially expressed genes were determined with adjusted p-values (q-values) in tumor samples (TP) according to normal samples (NT) of 55 LUAD patients by the limma-voom method using *limma* (Ritchie, 2015) and *edgeR* (McCarthy, 2012) R/Bioconductor packages. These packages can be used to determine differentially expressed genes in the RNA level for both microarray and RNAseq data. NCBI IDs and Hugo Symbols of the differentially expressed genes determined by the *biomaRt* package.

2.5. Active Subnetwork and Pathway Analysis

The active subnetworks of differentially expressed genes were determined using the R/Bioconductor *DEsubs* package (Vrahatis, 2016). The output of DEA containing differentially expressed genes with their Ensembl IDs and adjusted p-values (q-values) were used as input of *DEsubs* package. *DEsubs* package determines and represents the active subnetworks with their graphs both at subnetwork and pathway levels.

2.6. Statistical Analysis

Clinical data of patients were downloaded from the TCGA database using the *TCGABiolinks* package. Univariate and multivariate Cox Proportional Hazards Regression analysis (Cox, 1972) was performed for significant SNVs, CNVs, and DEGs from active subnetworks by using *survival* package (Therneau, 2000; Therneau, 2019). Patients were clustered based on the expression of DEGs from

identified subnetworks. Kaplan–Meier survival analysis (Kaplan & Meier, 1958) was performed for these clusters by using *survminer* package (Kassambara, 2019). The statistical significance of the results was determined using the log-rank test (Mantel, 1966). The significance level used for analysis was 0.05 for both p-value and q-value for SNV and Cox proportional analysis; 0.01 for DEA and CNV analysis.

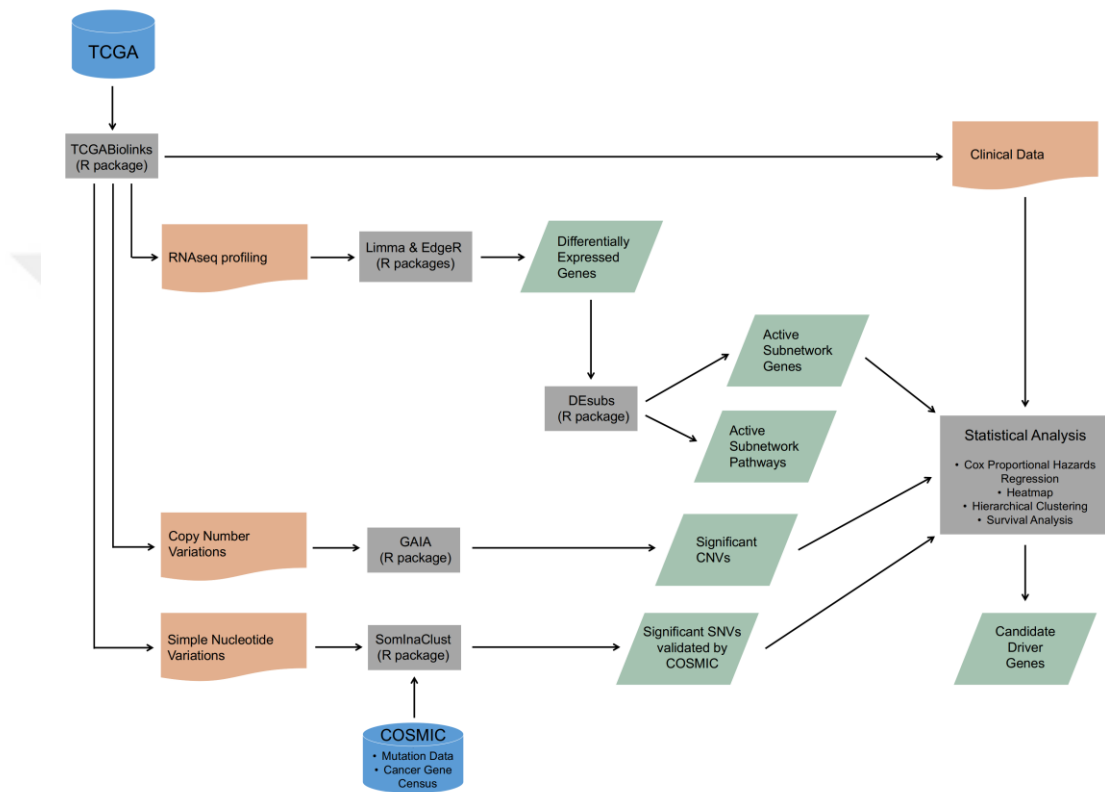


Figure 2.1. The flowchart of the pipeline

3. RESULTS

3.1. Identification of Significant Simple Nucleotide Variations

Mutation data of LUAD patients as maf file generated by mutect pipeline was downloaded by *TCGAbiolinks* package and *maftools* package was used to subset original maf file by tumor sample barcodes of 55 LUAD patients and 510 LUAD patients. Then, significant mutations for both 55 and 510 LUAD patients were determined separately with their roles as a tumor suppressor or oncogene by *SomInaClust* R package.

3.1.1. SNV analysis of 55 LUAD patients

Mutations are classified in nine variant classes (Missense mutation, Nonsense mutation, Frame-shift deletion, Frame-shift insertion, In-frame deletion, In-frame insertion, Translation start site, Splice site mutation and Non-stop mutation) in TCGA project. In total, 5819 genes are mutated in tumor samples of 55 LUAD patients and missense mutations are the most observed variant class, 8152 times with mean 148 per sample and median 120, followed by nonsense and frame-shift deletion. On the other hand, SNP (Single Nucleotide Polymorphism) is the most abundant variant type according to insertion and deletion. When we look at the SNV (Single Nucleotide Variation) class, C > A transversion and C > T transition are highly observed, followed by C > G transversion (Figure 3.1). The lower panel in Figure 3.1 summarizes the number of variants per sample, variant classification as boxplot and top ten mutated genes with their percentage of observance in tumor samples of 55 LUAD patients.

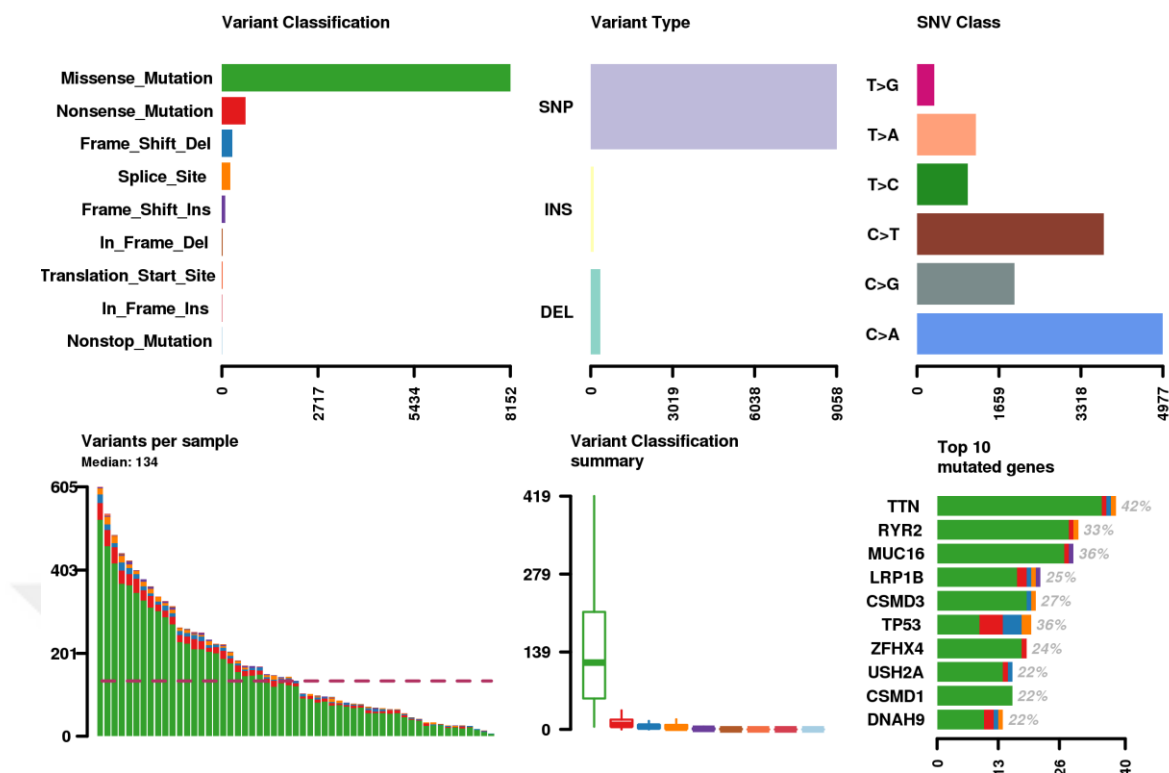


Figure 3.1. Summary plot for mutation data of 55 LUAD patients

TTN (Titin) and MUC16 (Mucin-16) genes are highly mutated genes, respectively in 42% and 36% of 49 of 55 tumor samples. The TP53 gene has remarkable mutation rate (36%), followed by RYR2 (33%), CSMD3 (27%), LRP1B (25%), ZFHX4 (24%), CSMD1 (22%), DNAH9 (22%) and USH2A (22%) (Figure 3.2). Although these genes have high mutation rates, these mutations may not be important. In order to determine important genes including significant mutations, we used another R package called *SomInaClust*. The high mutation rates of TTN and MUC16 may be caused by the length of these genes. TTN is the longest gene which consists of 311814 bp, and MUC16 has 168551 bp length in the human genome. These genes with high length are liable to a high number of mutations. Even though MUC16 has considerable mutations in ovarian cancer, the mutation rate in Lung cancer can be underestimated. The power of *SomInaClust* and other programs which detect driver genes is the elimination of unimportant genes although they have a high number of

mutations and the determination of important genes although they have a low number of mutations.

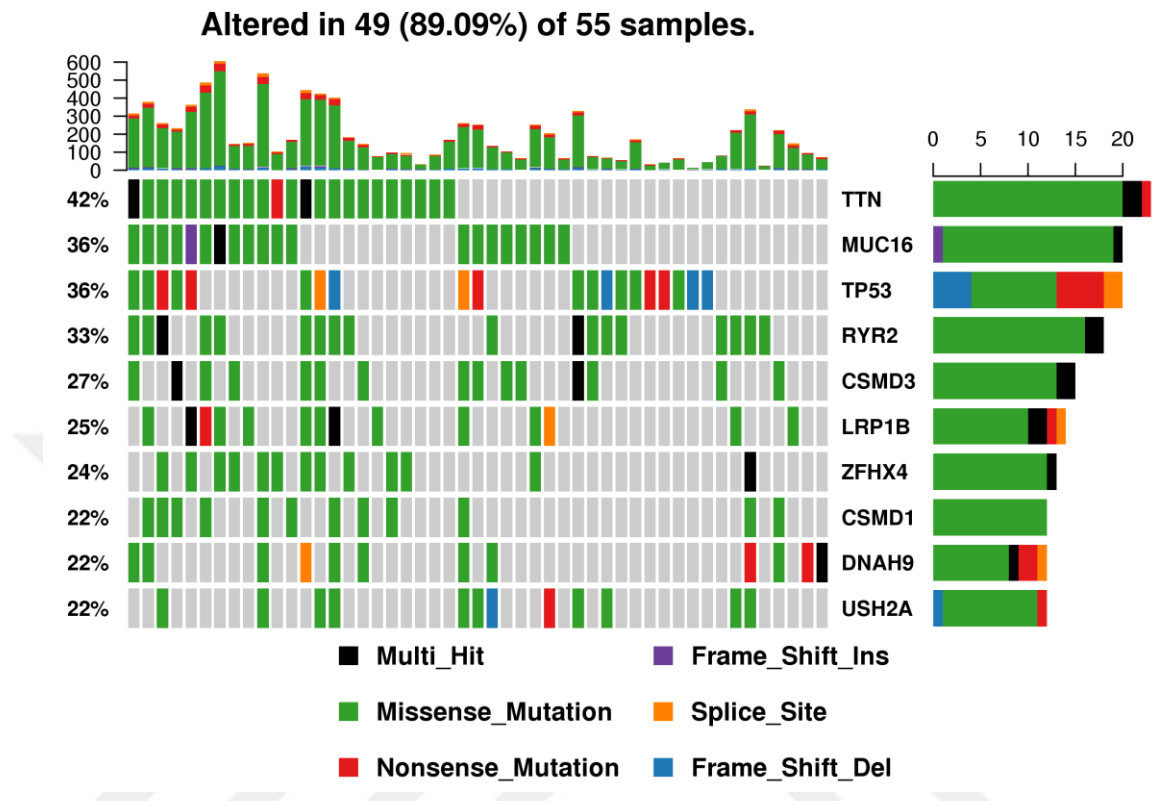


Figure 3.2. Oncoplot showing top ten mutated genes of tumor samples of 55 LUAD patients

Figure 3.3, shows the summary of mutation rates of substitution types calculated for each tumor sample. According to percentages of substitutions in samples, C > T transition has the highest median percentage followed by C > A transversion. In total, the transversion percentage is higher than the transition percentage.

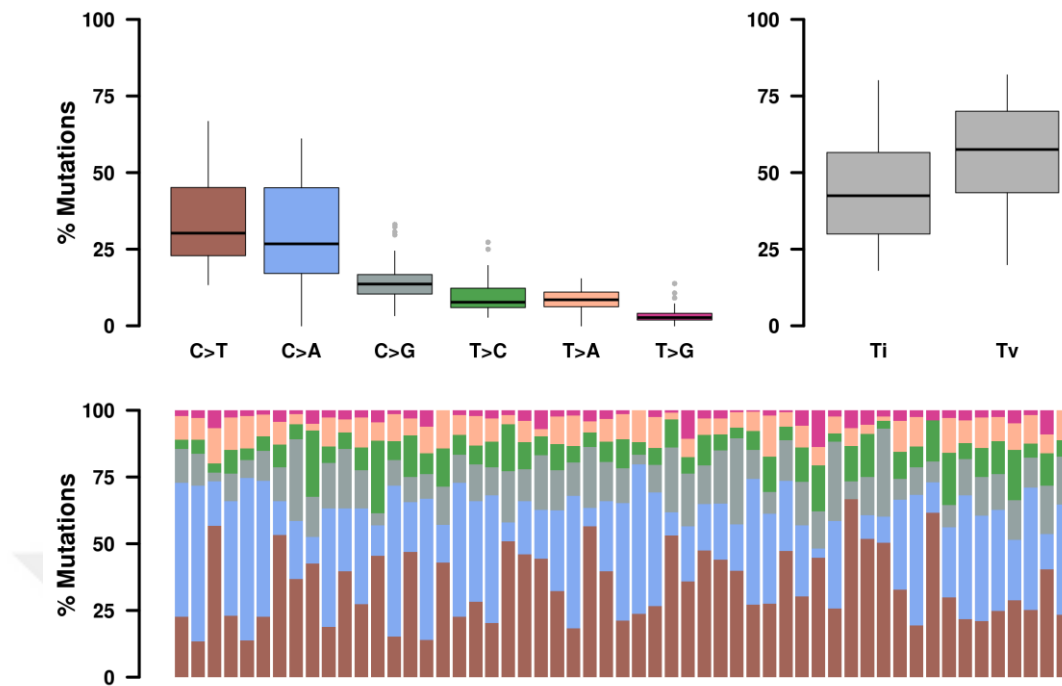


Figure 3.3. Transition and transversion rates in tumor samples of 55 LUAD patients

3.1.1.1. Determination of important mutated genes by SomInaClust package

In order to determine important genes which are mutated in tumor samples and may play a major role in Lung cancer, *SomInaClust* R package was used. *SomInaClust* determines candidate important mutated genes by using mutation patterns of them.

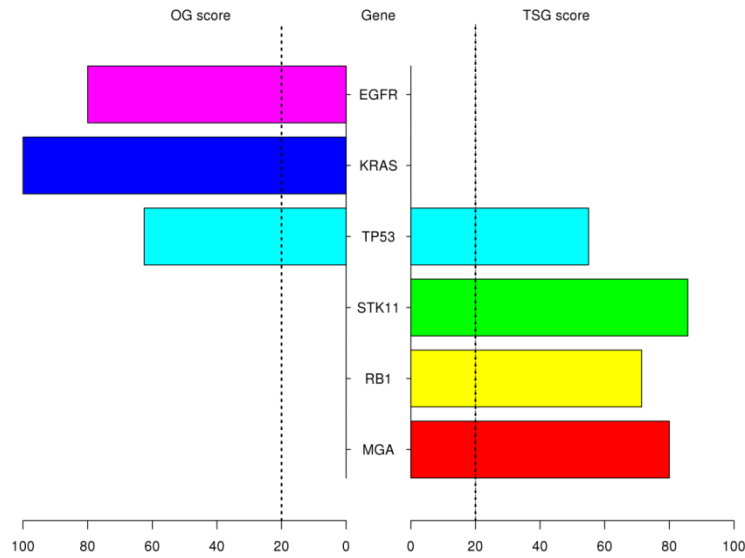


Figure 3.4. Pyramid plot of *SomInaClust* results in tumor samples of 55 LUAD patients

In Figure 3.4, EGFR, KRAS, TP53, STK11, RB1, and MGA genes are showed as candidate important genes in tumor samples of 55 LUAD patients. EGFR and KRAS genes are classified as oncogene and STK11, RB1 and MGA genes are classified as a tumor suppressor. But the TP53 gene has both tumor suppressor and oncogene scores. Table 3.1 summarizes the significant results of *SomInaClust*. EGFR, KRAS, TP53, STK11, and RB1 have highly significant estimation depending on reference information of cancer gene census. While EGFR and TP53 have a higher number of mutations, KRAS, STK11, RB1, and MGA have a lower number of mutations. Although TP53 gene has both OG score and TSG score, TP53 is classified as a tumor suppressor in Table 3.1.

Table 3.1. *SomInaClust* results showing significantly mutated genes in 55 LUAD patients

Gene	# Mutations	qDG	OG Score	TSG Score	Classification	CGC
EGFR	11	1.57e-12	80	0	OG	Dom
KRAS	8	1.57e-12	100	0	OG	Dom
TP53	20	4.8e-07	62.5	55	TSG	Rec
STK11	7	0.000106	0	85.7	TSG	Rec
RB1	7	0.0049	0	71.4	TSG	Rec
MGA	6	0.0217	0	80	TSG	NA

3.1.2. SNV analysis of 510 LUAD patients

In total, 16746 genes are mutated in tumor samples of 510 LUAD patients and missense mutations are the most observed variant class, 109672 times with mean 215 per sample and median 149, followed by nonsense and frame-shift deletion. SNP (Single Nucleotide Polymorphism) is the most abundant variant type against insertion and deletion. C > A transversion and C > T transition are highly observed SNV class, followed by C > G transversion (Figure 3.5). The mutation pattern of 510 LUAD patients is similar to the pattern of 55 LUAD patients except for the number of mutated genes and variant classification.

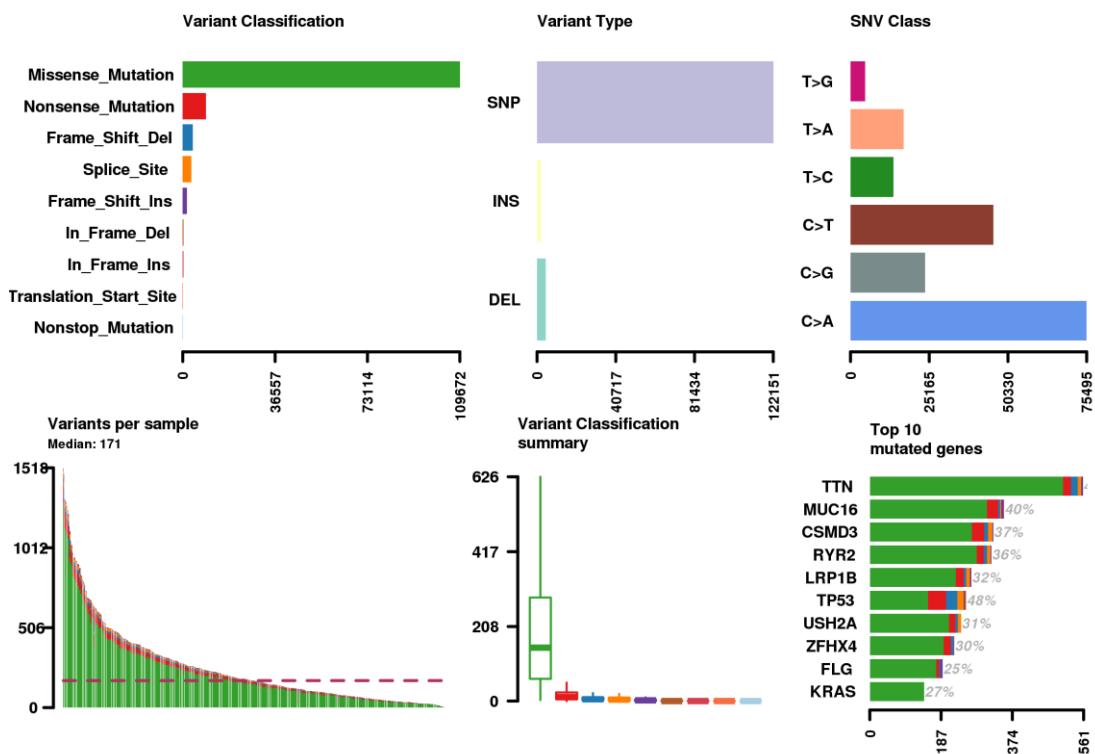


Figure 3.5. Summary plot for mutation data of 510 patients in LUAD project

TP53 gene is highly mutated followed by TTN (Titin) and MUC16 (Mucin-16) genes, respectively in 48%, 45% and 40% of 452 of 509 tumor samples. The other highly mutated genes are CSMD3 (37%), RYR2 (36%), LRP1B (32%), USH2A (31%), ZFHX4 (30%), KRAS (27%) and FLG (25%) (Figure 3.6). TP53, TTN, MUC16,

CSMD3, RYR2, LRP1B, USH2A, ZFHX4 except KRAS and FLG are in the list of top ten mutated genes in tumor samples of both 55 and 510 LUAD patients.

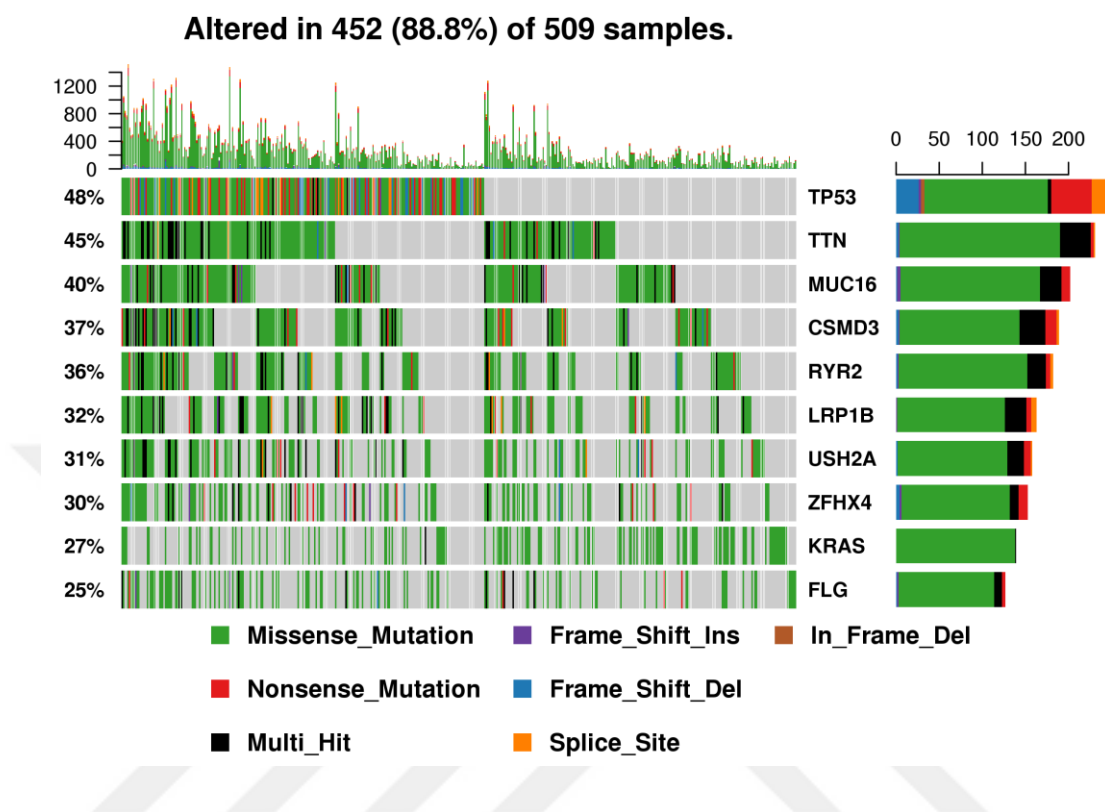


Figure 3.6. Oncoplot showing top ten mutated genes of tumor samples of 510 LUAD patients

Figure 3.7, shows the summary of mutation rates of substitution types calculated for each tumor sample. According to percentages of substitutions in samples, C > A transversion has the highest median percentage followed by C > T transition and C > G transversion. In total, the transversion percentage is higher than the transition percentage for also 510 LUAD patients.

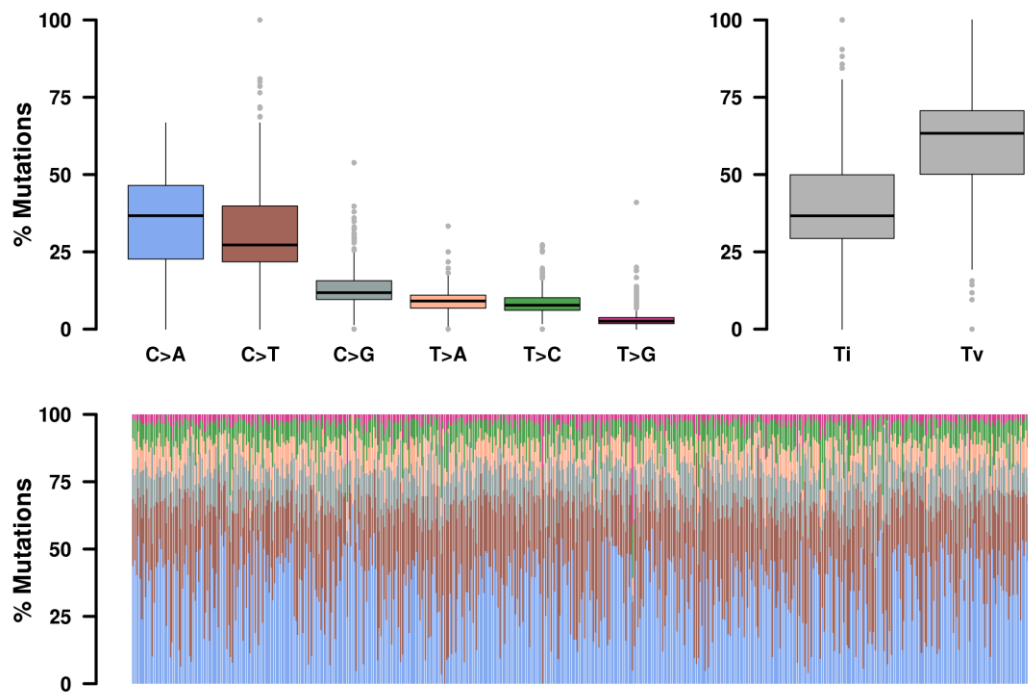


Figure 3.7. Transition and transversion rates in tumor samples of 510 LUAD patients

3.1.2.1. Determination of important mutated genes by SomInaClust package

In Figure 3.8, the top 40 genes are shown as candidate important genes in tumor samples of 510 LUAD patients. EGFR, KRAS, TP53, STK11, RB1 and MGA genes are also in the list of SomInaClust results with other 76 significant mutated genes including BRAF, PIK3CA, ATM, ERBB2, MAP2K1, SMAD4 and CDKN2A which are related with lung cancer and listed in cancer gene census.

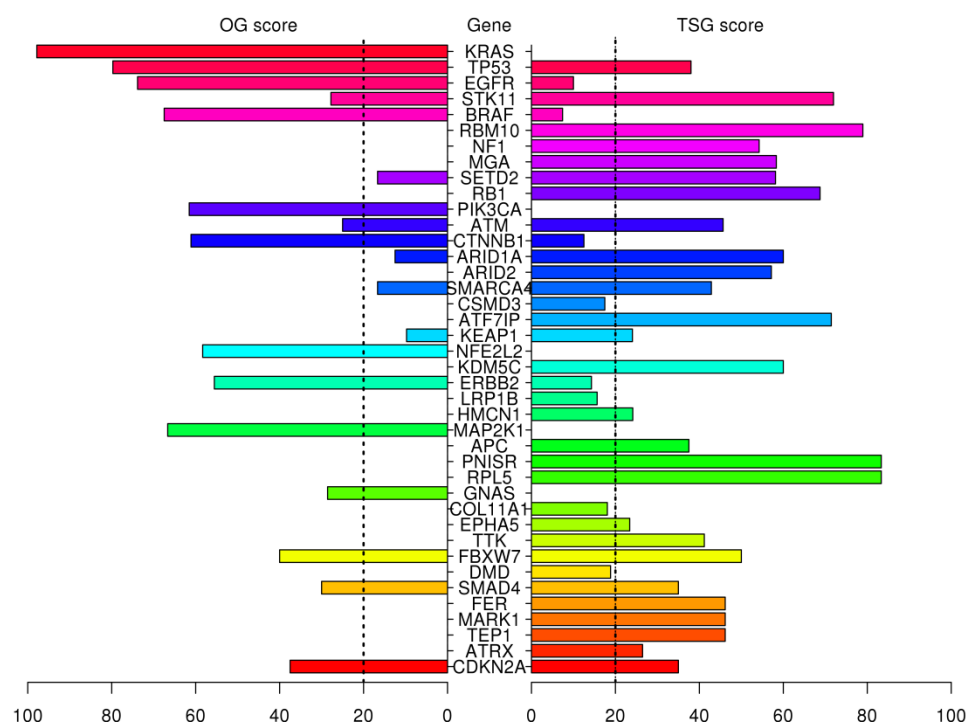


Figure 3.8. Pyramid plot of *SomInaClust* results in tumor samples of 510 LUAD patients

Table 3.2, summarizes the *SomInaClust* results showing 82 significant mutated genes, most of which are classified as oncogene (OG) or tumor suppressor gene (TSG) while some genes, CSMD3, COL11A1, DMD, SPTA1, USH2A, CDH10, and DST cannot be classified as oncogene or tumor suppressor because their scores are under threshold. LRP1B and KMT2C cannot also be classified although they are listed as a tumor suppressor in cancer gene census. The known cancer related genes, KRAS, TP53, EGFR, STK11, BRAF, NF1, SETD2, RB1, PIK3CA, ATM, CTNNB1, ARID1A, ARID2, SMARCA4, NFE2L2, KDM5C, ERBB2, MAP2K1, APC, GNAS, FBXW7, SMAD4, ATRX, CDKN2A, CDKN1B, PBRM1, AKT1, RAF1, MAP2K4, B2M, BAP1 are classified as listed in cancer gene census. But the classification of RPL5 and IDH1 has a contradiction with cancer gene census. On the other hand, *SomInaClust* determined and classified the genes such as RBM10, MGA, ATF7IP, KEAP1, HMCN1, PNISR, EPHA5, TTK, FER, MARK1, TEP1, MYO9A, ZNF800, CMTR2, RASA1, DHX15, IQGAP2, LTN1, SMARCA1, FHOD3, CPVL,

MAP3K12, TOP2B, ROCK1, AKAP6, SENP1, SP1, WISP3, RAD50, COL28A1, SCAF8, STK31, YLPM1, IQUB, MARK2, NAA15, VPS13C, ZBBX, DGKB, FBN2 which are not listed in cancer gene census (Table 3.2). SomInaClust could determine candidate driver genes such as CDKN1B, SENP1, SP1, WISP3, and AKT1 genes although they have less than 5 mutations.

Table 3.2. SomInaClust results showing significantly mutated genes in 510 LUAD patients

Gene	# Mutations	qDG	OG Score	TSG Score	Classification	CGC
KRAS	143	1.97e-250	97.8	0	OG	Dom
TP53	253	2.52e-135	79.7	38	TSG	Rec
EGFR	73	8.97e-84	73.8	10	OG	Dom
STK11	83	4.6e-61	27.8	72	TSG	Rec
BRAF	44	8.07e-51	67.5	7.4	OG	Dom
RBM10	39	9.06e-31	0	78.9	TSG	NA
NF1	63	5.37e-25	0	54.2	TSG	Rec
MGA	52	6.46e-23	0	58.3	TSG	NA
SETD2	44	1.34e-20	16.7	58.1	TSG	Rec
RB1	32	4.99e-20	0	68.8	TSG	Rec
PIK3CA	27	1.36e-19	61.5	0	OG	Dom
ATM	48	5.18e-18	25	45.7	TSG	Rec
CTNNB1	21	3.32e-15	61.1	12.5	OG	Dom
ARID1A	30	1.76e-14	12.5	60	TSG	Rec
ARID2	29	2.83e-12	0	57.1	TSG	Rec
SMARCA4	48	2.23e-11	16.7	42.9	TSG	Rec
CSMD3	324	6.25e-10	0	17.5	NA	NA
ATF7IP	17	1.84e-08	0	71.4	TSG	NA
KEAP1	90	1.91e-08	9.8	24.1	TSG	NA
NFE2L2	14	2.83e-07	58.3	0	OG	Dom
KDM5C	16	1.76e-06	0	60	TSG	Rec
ERBB2	13	6.94e-06	55.6	14.3	OG	Dom
LRP1B	267	6.04e-05	0	15.6	NA	Rec
HMCN1	97	8.93e-05	0	24.1	TSG	NA
MAP2K1	9	0.000263	66.7	0	OG	Dom
APC	24	0.000272	0	37.5	TSG	Rec
PNISR	6	0.000626	0	83.3	TSG	NA
RPL5	7	0.000626	0	83.3	TSG	Dom
GNAS	19	0.000962	28.6	0	OG	Dom
COL11A1	129	0.00139	0	18.1	NA	NA
EPHA5	66	0.00221	0	23.4	TSG	NA

Table 3.2. (continue)

Gene	# Mutations	qDG	OG Score	TSG Score	Classification	CGC
TTK	18	0.00221	0	41.2	TSG	NA
FBXW7	12	0.0028	40	50	TSG	Rec
DMD	99	0.00349	0	18.8	NA	NA
SMAD4	20	0.00379	30	35	TSG	Rec
FER	16	0.0043	0	46.2	TSG	NA
MARK1	21	0.0043	0	46.2	TSG	NA
TEP1	29	0.0043	0	46.2	TSG	NA
ATRX	35	0.00463	0	26.5	TSG	Rec
CDKN2A	21	0.00585	37.5	35	TSG	Rec
MYO9A	19	0.00615	0	42.9	TSG	NA
ZNF800	17	0.00615	0	42.9	TSG	NA
CMTR2	26	0.00674	0	55.6	TSG	NA
RASA1	9	0.00674	0	55.6	TSG	NA
CDKN1B	5	0.00674	0	80	TSG	Rec
DHX15	7	0.00674	0	80	TSG	NA
IQGAP2	28	0.00816	0	40	TSG	NA
LTN1	19	0.00816	0	40	TSG	NA
SMARCA1	19	0.00816	0	40	TSG	NA
SPTA1	164	0.00971	0	17.6	NA	NA
FHOD3	31	0.0122	0	30.4	TSG	NA
CPVL	8	0.0161	0	66.7	TSG	NA
MAP3K12	8	0.0161	0	66.7	TSG	NA
TOP2B	9	0.0161	0	66.7	TSG	NA
ROCK1	21	0.0163	0	35.3	TSG	NA
PBRM1	12	0.0172	0	45.5	TSG	Rec
AKAP6	40	0.0195	0	28	TSG	NA
SENPI	3	0.0241	0	100	TSG	NA
SP1	4	0.0241	0	100	TSG	NA
WISP3	4	0.0241	0	100	TSG	NA
RAD50	13	0.0243	20	41.7	TSG	NA
COL28A1	19	0.0243	0	41.7	TSG	NA
SCAF8	18	0.0243	0	41.7	TSG	NA
STK31	19	0.0243	0	41.7	TSG	NA
IDH1	6	0.0248	40	25	TSG	Dom
USH2A	240	0.0263	0	13.2	NA	NA
YLPM1	23	0.0269	0	31.6	TSG	NA
IQUB	12	0.0272	0	57.1	TSG	NA
MARK2	10	0.0272	0	57.1	TSG	NA
NAA15	8	0.0272	0	57.1	TSG	NA

Table 3.2. (continue)

Gene	# Mutations	qDG	OG Score	TSG Score	Classification	CGC
CDH10	99	0.028	0	16.4	NA	NA
AKT1	3	0.0296	66.7	0	OG	Dom
RAF1	7	0.031	66.7	0	OG	Dom
VPS13C	39	0.0332	0	25	TSG	NA
ZBBX	28	0.0333	0	30	TSG	NA
DST	67	0.0333	0	19.1	NA	NA
KMT2C	52	0.0388	0	18.8	NA	Rec
DGKB	38	0.0431	0	28.6	TSG	NA
MAP2K4	8	0.045	33.3	50	TSG	Rec
FBN2	93	0.045	0	20.5	TSG	NA
B2M	8	0.045	0	50	TSG	Rec
BAP1	8	0.045	0	50	TSG	Rec

3.2. Identification of Significant Copy Number Variations

CNVs (Copy Number Variations) are also important aberrations which result in alterations in gene expression in tumorigenesis and tumor growth. In order to determine significant CNVs among tumor samples of 55 and 510 LUAD patients, gaia package was used.

3.2.1. CNV analysis of 55 LUAD patients

Significant recurrent CNVs in tumor samples of 55 LUAD patients, over the orange q value thresholds (0.01), are mostly observed on Chromosome 1, 8, 9, and 17. But Chromosome 1 has the highest number of amplification followed by Chromosome 8. Chromosome 9 has the highest number of deletions followed by Chromosome 17 as seen in Figure 3.9.

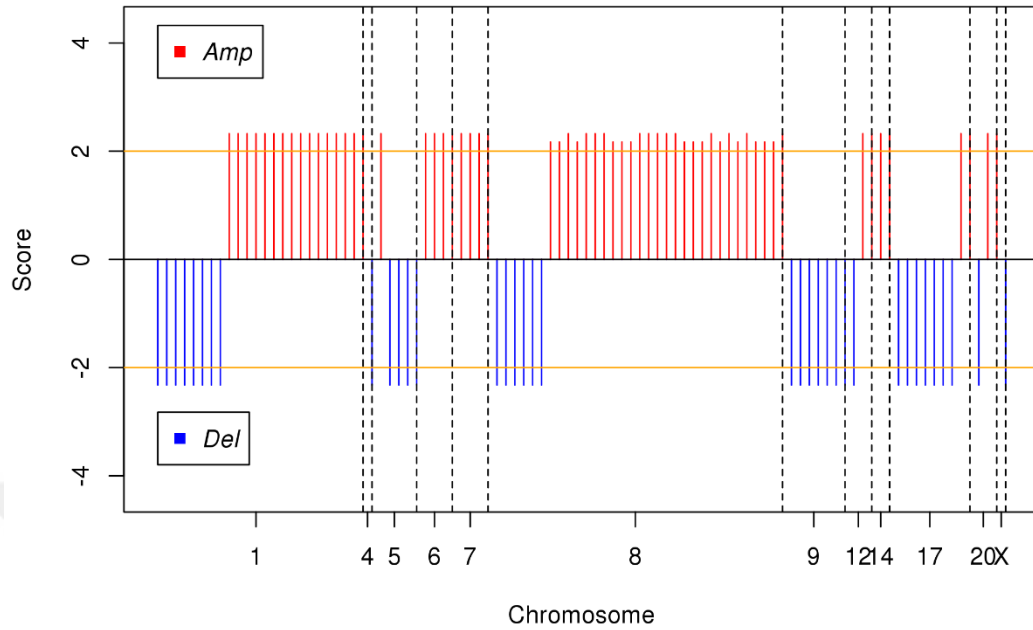


Figure 3.9. Recurrent CNVs on all chromosomes in tumor samples of 55 LUAD patients

Top ten significant amplified and deleted genes which are all from chromosome 1 are shown in Table 3.3. Chromosome 1 has the highest number of gene aberration with 2006 amplified or deleted genes followed by Chromosome 8 with 1029 aberrant genes and Chromosome 19 with 785 aberrant genes.

Table 3.3. Top ten significant deleted and amplified genes in 55 LUAD patients

Gene Symbol	Aberration	q-value	Aberrant Region	Gene Region
RN7SKP285	Del	0.00474651	1:103501576-107318961	1:103523562-103523879
RNPC3	Del	0.00474651	1:103501576-107318961	1:103525691-103555239
AMY2B	Del	0.00474651	1:103501576-107318961	1:103553815-103579534
ACTG1P4	Del	0.00474651	1:103501576-107318961	1:103569553-103570674
AMY2A	Del	0.00474651	1:103501576-107318961	1:103616811-103625780
AMY1A	Del	0.00474651	1:103501576-107318961	1:103655290-103664554
AC105272.1	Del	0.00474651	1:103501576-107318961	1:103668071-103668268
AMY1B	Del	0.00474651	1:103501576-107318961	1:103687415-103696680
AMYP1	Del	0.00474651	1:103501576-107318961	1:103713723-103719871
AMY1C	Del	0.00474651	1:103501576-107318961	1:103750406-103758690

Table 3.3. (continue)

Gene Symbol	Aberration	q-value	Aberrant Region	Gene Region
PLEKHO1	Amp	0.00474651	1:150131878-150768299	1:150149183-150164720
AC242988.2	Amp	0.00474651	1:150131878-150768299	1:150173049-150181429
RN7SL480P	Amp	0.00474651	1:150131878-150768299	1:150211632-150211925
ANP32E	Amp	0.00474651	1:150131878-150768299	1:150218417-150236156
RNU2-17P	Amp	0.00474651	1:150131878-150768299	1:150236967-150237156
AC242988.1	Amp	0.00474651	1:150131878-150768299	1:150255095-150257286
CA14	Amp	0.00474651	1:150131878-150768299	1:150257251-150265078
APH1A	Amp	0.00474651	1:150131878-150768299	1:150265399-150269580
C1orf54	Amp	0.00474651	1:150131878-150768299	1:150268200-150280916
CIART	Amp	0.00474651	1:150131878-150768299	1:150282543-150287093

When we look at the pathways of amplified and deleted genes, only Small Cell Lung Cancer Signaling amplification is directly related to lung cancer. On the other hand, amplification in immune system pathways such as Complement System, Dendritic Cell Maturation and Natural Killer Cell Signaling recall the immune system dysregulation in cancer to mind (Figure 3.10). Role of p14/p19ARF in Tumor Suppression pathway is the most important pathway of deleted genes. The other possible important pathways are Calcium Signaling, Tight Junction Signaling and Actin Cytoskeleton Signaling which may play a role in apoptosis and metastasis events (Figure 3.11).

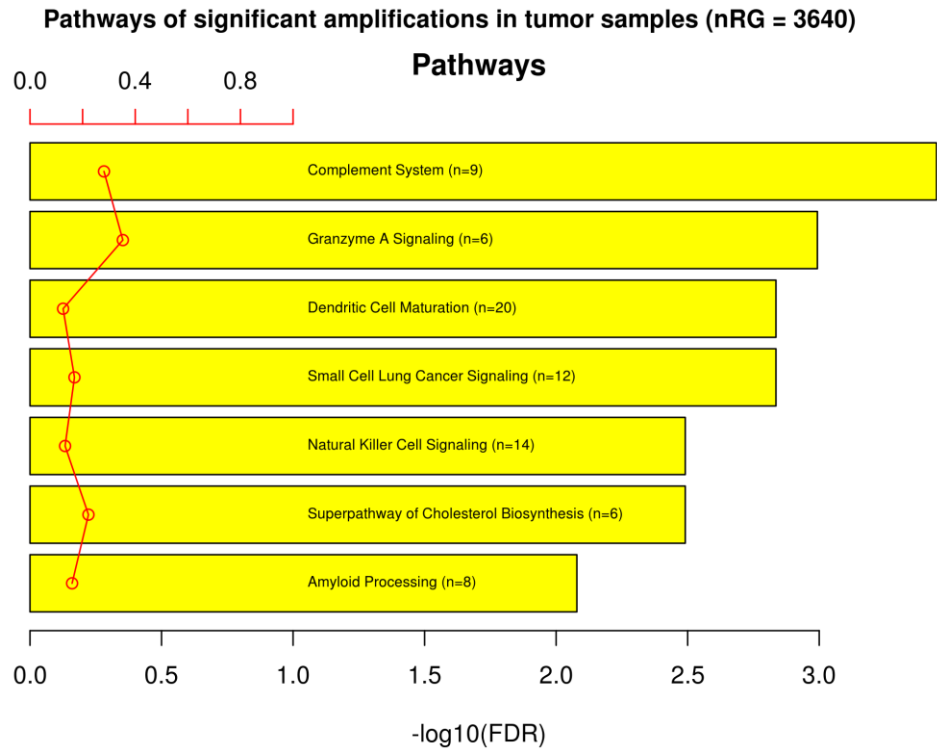


Figure 3.10. Pathways of significant amplification in tumor samples of 55 LUAD patients

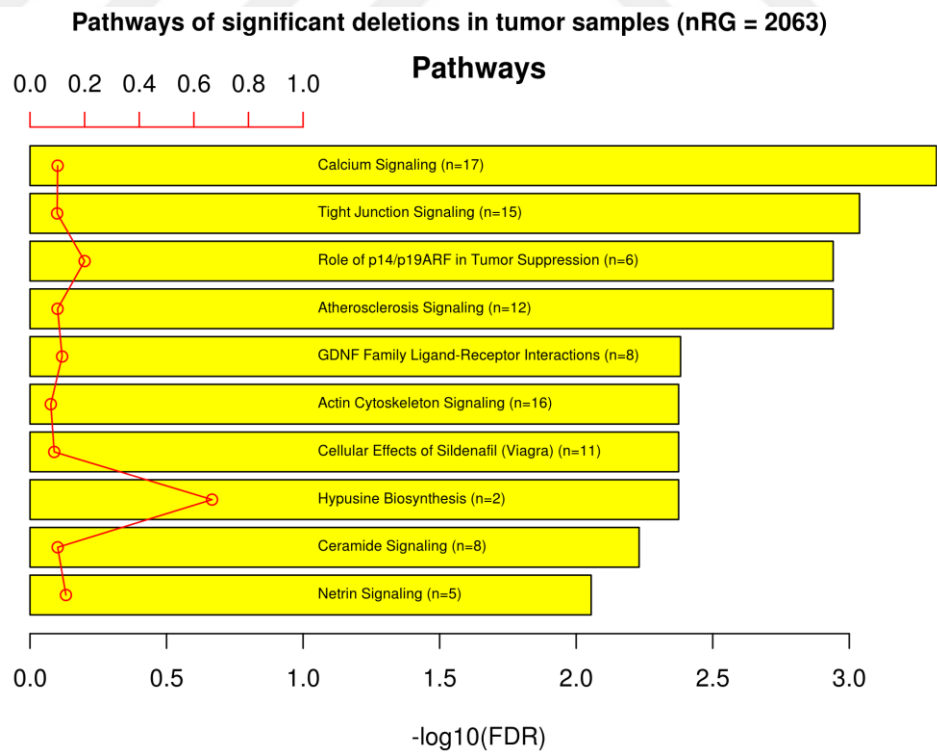


Figure 3.11. Pathways of significant deletions in tumor samples of 55 LUAD patients

3.2.2. CNV analysis of 510 LUAD patients

Significant recurrent CNVs in tumor samples of 510 LUAD patients, over the orange q value thresholds (0.05), are mostly observed on Chromosome 4, 9, 10, 11, 12, 13, 14, 16, 18 and 20. But Chromosome 11 has the highest number of aberrations followed by Chromosome 9, 16 and 18. Chromosome 4, 9, 10, 12 and 16 have mostly amplification (Figure 3.12). The pattern of CNVs in tumor samples of 510 LUAD patients has a marked difference from the CNV pattern in tumor samples of 55 LUAD patients (Figure 3.9).

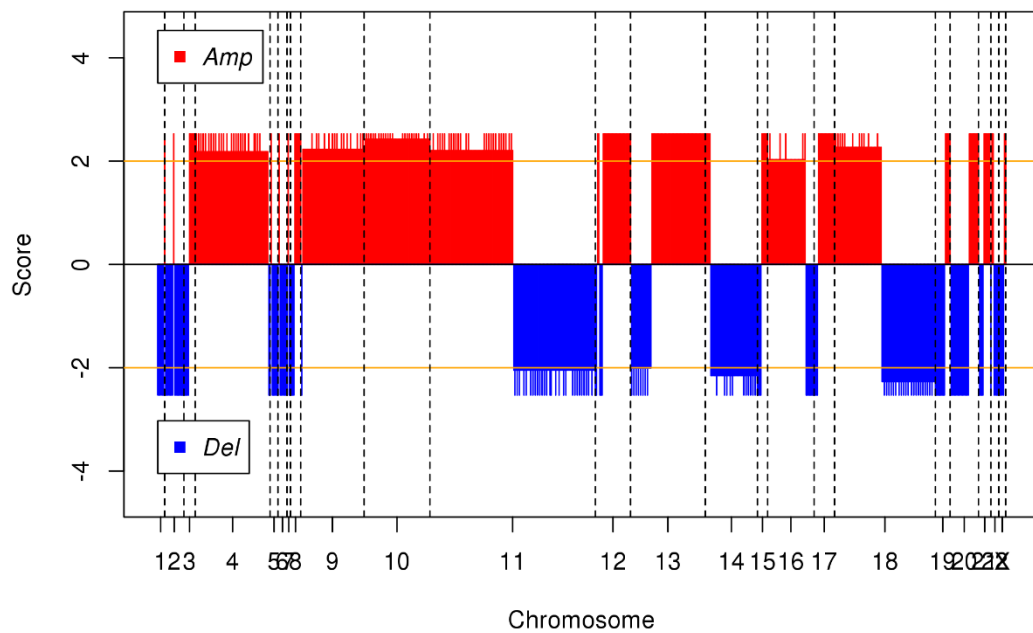


Figure 3.12. Recurrent CNVs on all chromosomes in tumor samples of 510 LUAD patients

Top ten significant amplified and deleted genes which are all from chromosome 1 are shown in Table 3.4. Chromosome 1 has the highest number of gene aberration with 3124 amplified or deleted genes followed by Chromosome 6 with 2911 aberrant genes and Chromosome 3 with 2149 aberrant genes.

Table 3.4. Top ten significant amplified and deleted genes in 510 LUAD patients

Gene Symbol	Aberration	q-value	Aberrant Region	Gene Region
AL359821.1	Del	0.0029609	1:71621685-71778398	1:71738173-71738354
GDI2P2	Del	0.0029609	1:71928758-119984738	1:72274552-72275159
AL513166.2	Del	0.0029609	1:71928758-119984738	1:72283170-72753772
RPL31P12	Del	0.0029609	1:71928758-119984738	1:72301472-72301829
AL583808.1	Del	0.0029609	1:71928758-119984738	1:72636547-72899240
RNU6-1246P	Del	0.0029609	1:71928758-119984738	1:72717663-72717769
AL583808.2	Del	0.0029609	1:71928758-119984738	1:72765031-72791282
AL583808.3	Del	0.0029609	1:71928758-119984738	1:72793104-72854475
AL732618.1	Del	0.0029609	1:71928758-119984738	1:72979014-72979314
KRT8P21	Del	0.0029609	1:71928758-119984738	1:73104792-73106282
SF3B4	Amp	0.0029609	1:149907993-247650984	1:149923317-149927803
MTMR11	Amp	0.0029609	1:149907993-247650984	1:149928651-149936879
OTUD7B	Amp	0.0029609	1:149907993-247650984	1:149937812-150010726
AC244033.2	Amp	0.0029609	1:149907993-247650984	1:150045660-150067701
AC244033.1	Amp	0.0029609	1:149907993-247650984	1:150053864-150055034
VPS45	Amp	0.0029609	1:149907993-247650984	1:150067279-150145329
PLEKHO1	Amp	0.0029609	1:149907993-247650984	1:150149183-150164720
AC242988.2	Amp	0.0029609	1:149907993-247650984	1:150173049-150181429
RN7SL480P	Amp	0.0029609	1:149907993-247650984	1:150211632-150211925
ANP32E	Amp	0.0029609	1:149907993-247650984	1:150218417-150236156

When we look at the pathways of aberrant genes, remarkable pathways are NF- κ B Signaling and Molecular Mechanisms of Cancer pathways which are directly related to cancer (Figure 3.13). But most of the altered pathways are immune system pathways as the pathways which have deleted and amplified genes (Figure 3.14, 3.15).

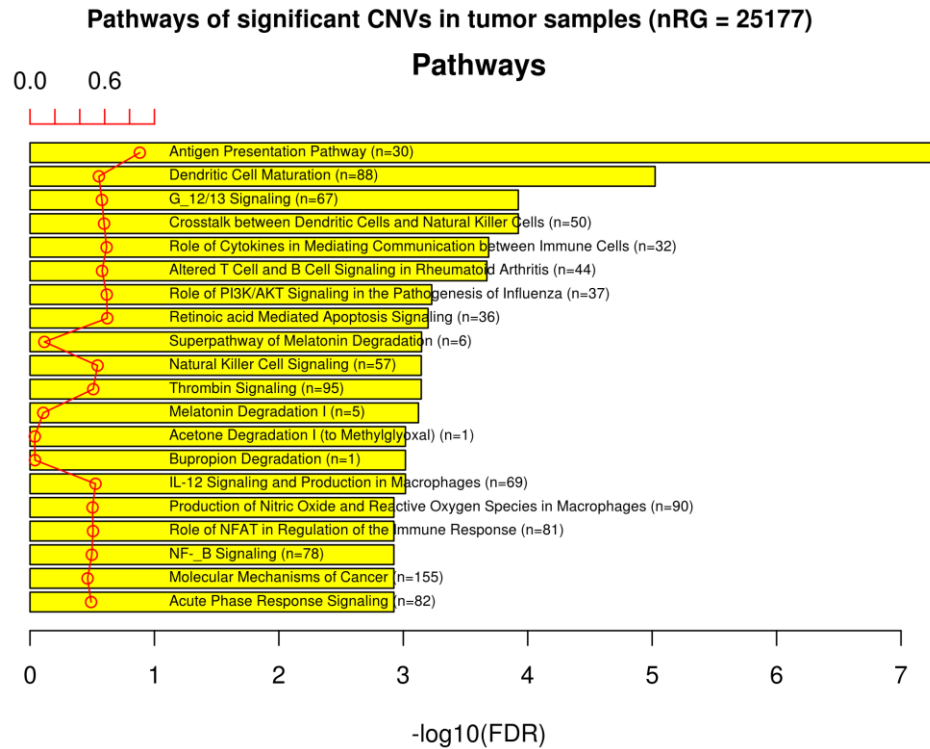


Figure 3.13. Top 20 Pathways of significant aberrations in tumor samples of 510 LUAD patients

Vast majority of the pathways which have amplified and deleted genes are immune system pathways except for Type I Diabetes Mellitus Signaling pathway including 41 amplified genes (Figure 3.14); G α 12/13 Signaling, DNA Methylation and Transcriptional Repression Signaling, Androgen Signaling, p70S6K Signaling and Regulation of Epithelial-Mesenchymal Transition Pathway which have deleted genes (Figure 3.15). Overwhelming aberrations in immune system pathways suggests that immune system dysregulation and tumor-immune system relationship are the key mechanisms in tumorigenesis.

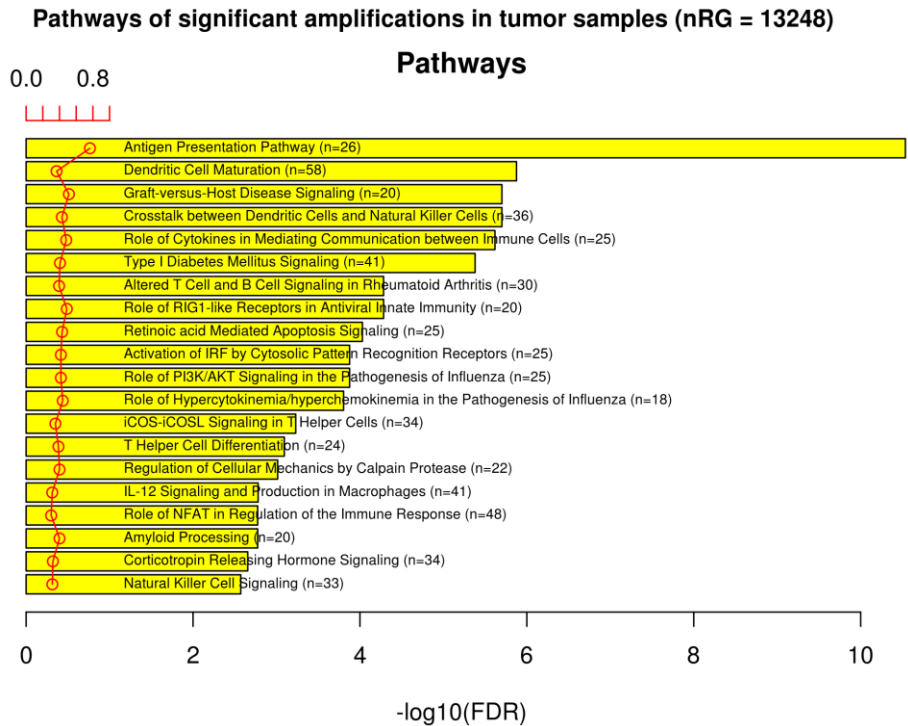


Figure 3.14. Top 20 pathways of significant amplification in tumor samples of 510 LUAD patients

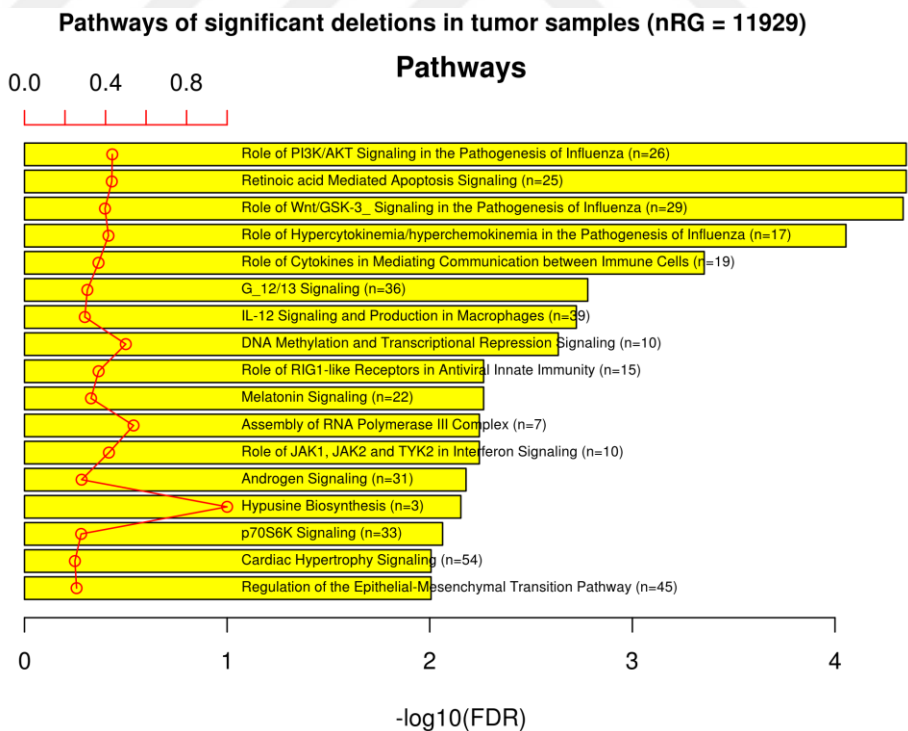


Figure 3.15. Top 20 Pathways of significant deletions in tumor samples of 510 LUAD patients

3.3. Differential Expression Analysis (DEA)

The Transcriptome Profiling data of LUAD patients in mRNA expression level (as unnormalized HTSeq raw counts), generated by TCGA RNAseq pipeline was downloaded by TCGABiolinks R package and prepared for analysis. Differentially expressed genes were determined with adjusted p-values (q-values) in tumor samples (TP) of 55 and 510 LUAD patients according to normal samples (NT) of 55 LUAD patients by the limma-voom method using limma and edgeR packages.

3.3.1. DEA of 55 LUAD patients

The volcano plot in Figure 3.16, shows the differentially expressed genes (DEGs) as dots of which black ones represent the genes which have differential expression less than two-fold and not significant while red ones represent upregulated and green ones downregulated more than two-fold ($\log_2=1$) significantly (FDR corrected p-value < 0.01). As a result of this analysis, 3575 genes are dysregulated more than two-fold with 0.01 FDR significance.

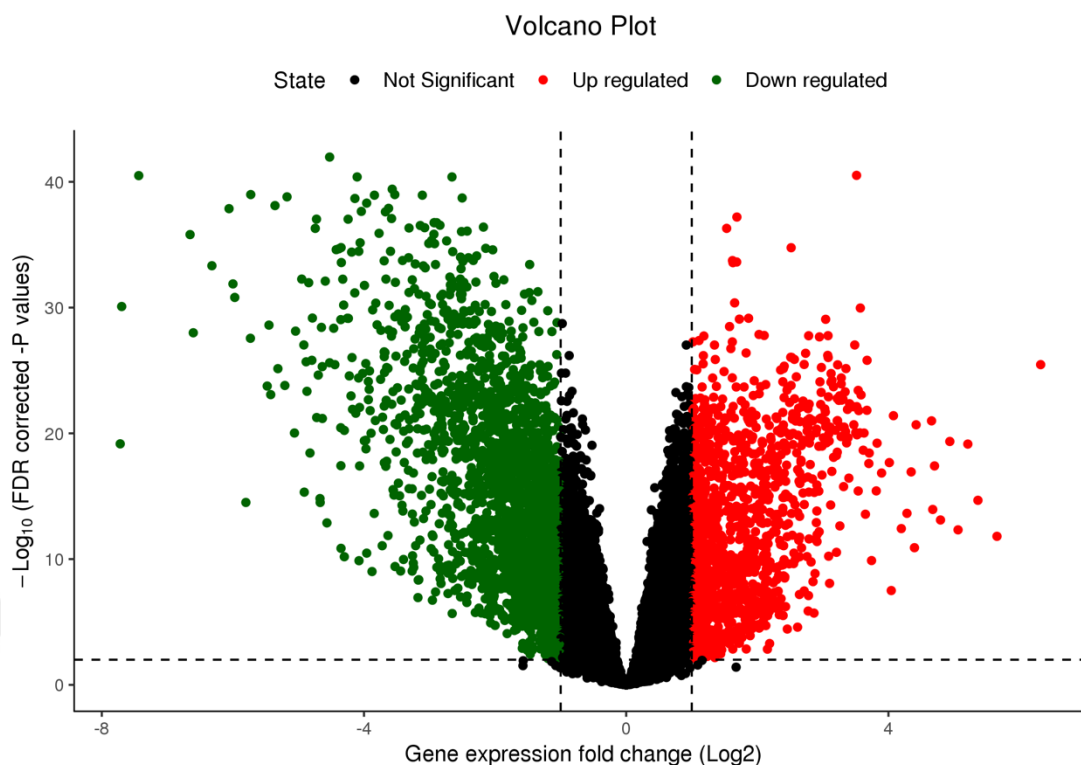


Figure 3.16. Volcano plot of differentially expressed genes in tumor samples of 55 LUAD patients

As a result of DEA, differentially expressed genes (DEGs) are determined with their log Fold Change (logFC), adjusted p-value (q-value), Entrez gene IDs and HGNC symbols after enrichment analysis. The top 10 down-regulated and up-regulated genes are shown in Table 3.5 and Table 3.6. The list of DEGs was used for pathway analysis and active subnetwork analysis.

Table 3.5. Top ten significant down-regulated genes in tumor samples of 55 LUAD patients

ensembl_gene_id	entrezgene	hgnc_symbol	logFC	adj.P.Val
ENSG00000182010	219790	RTKN2	-4.52455117194123	1.07397390772473e-42
ENSG00000158764	142683	ITLN2	-7.4364942528429	3.19924465283634e-41
ENSG00000102683	6445	SGCG	-4.10485571819757	4.07515928515459e-41
ENSG00000198873	2869	GRK5	-2.65790712992412	4.07515928515459e-41
ENSG00000107742	9806	SPOCK2	-3.56967403596283	3.85300139768808e-40
ENSG00000170323	2167	FABP4	-5.72790493543673	1.03033381509032e-39
ENSG00000135063	9413	FAM189A2	-3.53046742312343	1.03117504787973e-39
ENSG00000186994	256949	KANK3	-3.1101996380779	1.15468325581686e-39
ENSG00000150625	2823	GPM6A	-5.17438700689996	1.5648953870669e-39
ENSG00000154721	58494	JAM2	-2.50261146610761	1.92231892168565e-39

Table 3.6. Top ten significant up-regulated genes in tumor samples of 55 LUAD patients

ensembl_gene_id	entrezgene	hgnc_symbol	logFC	adj.P.Val
ENSG00000183010	5831	PYCR1	3.5139225242735	3.06017765569688e-41
ENSG00000059573	5832	ALDH18A1	1.68852856318992	6.30895314373162e-38
ENSG00000164466	94081	SFXN1	1.5322079314688	5.01920971916517e-37
ENSG00000135052	51280	GOLM1	2.51608337184892	1.73125209540521e-35
ENSG00000180198	1104	RCC1	1.62119814668367	1.82637777402036e-34
ENSG00000155660	9601	PDIA4	1.6848754492746	2.37855372052335e-34
ENSG00000096063	6732	SRPK1	1.62823462104507	2.66740561460568e-34
ENSG00000128050	10606	PAICS	1.65390171937903	4.22169230063646e-31
ENSG00000111344	8437	RASAL1	3.57173273242386	1.08251787193746e-30
ENSG00000173457	26472	PPP1R14B	1.86684316566064	7.07845976872399e-30

The most significant dysregulated pathways of DEGs are including Inhibition of Matrix Metalloproteases, EIF2 Signaling, Cell Cycle: G2/M DNA Damage Checkpoint Regulation, cAMP-mediated Signaling and Cell Cycle Control of Chromosomal Replication, which are related to cancer (Figure 3.17). But there are also other pathways which may not be related to cancer directly. In order to determine dysregulated pathways which are important and highly related to cancer, we can perform enrichment analysis separately as downregulated and upregulated pathways.

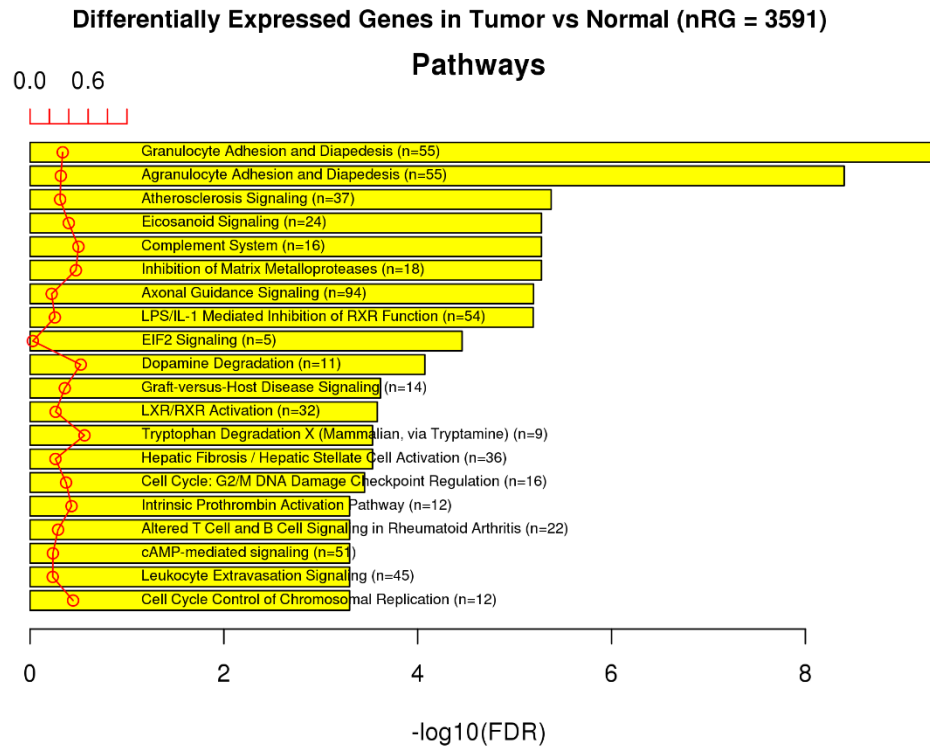


Figure 3.17. Top 20 dysregulated pathways of gene expression in tumor samples of 55 LUAD patients

As we see in Figure 3.18, most significantly upregulated pathways are Cell Cycle, Mitosis, DNA Damage and Metabolic pathways such as Glycolysis and HIF1 Signaling which are expected to be upregulated in tumor cells. And the other pathways such as Granulocyte adhesion and diapedesis, GADD45 signaling, Colanic acid building blocks biosynthesis, and Agranulocyte adhesion and diapedesis pathways need to be considered for cancer relation.

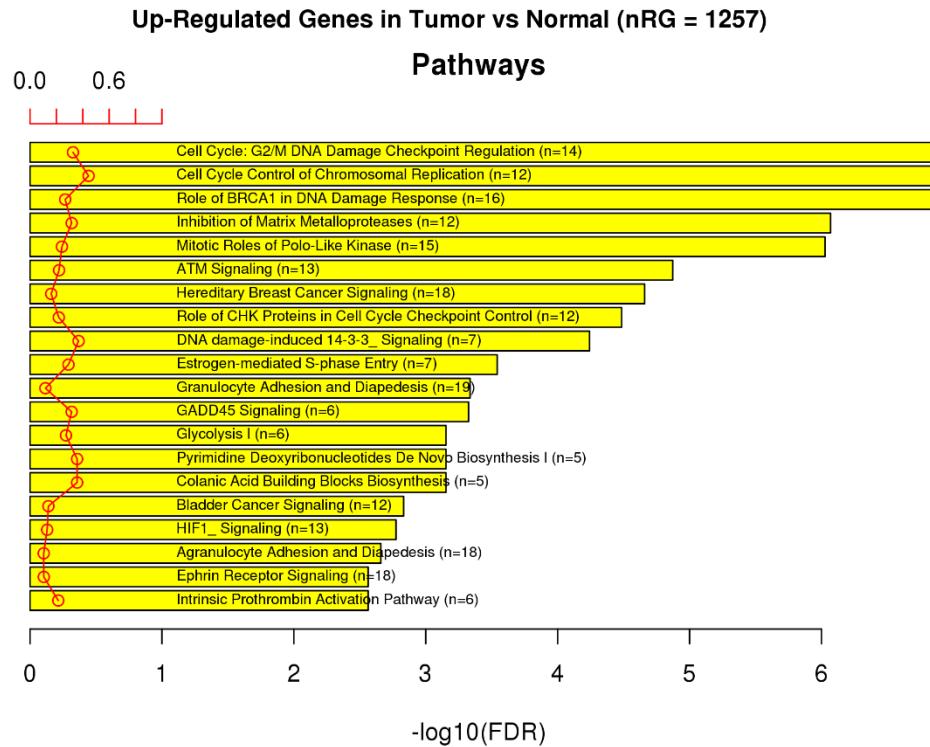


Figure 3.18. Upregulated pathways of gene expression in tumor samples of 55 LUAD patients

When we look at the pathways consisting downregulated genes, in Figure 3.19, the most attention-grabbing downregulated pathways are cAMP-mediated signaling, G-protein coupled receptor signaling, Gai signaling and other immune system pathways such as Complement system, Granulocyte/Agranulocyte adhesion and diapedesis, Dendritic cell maturation and T Helper cell differentiation.

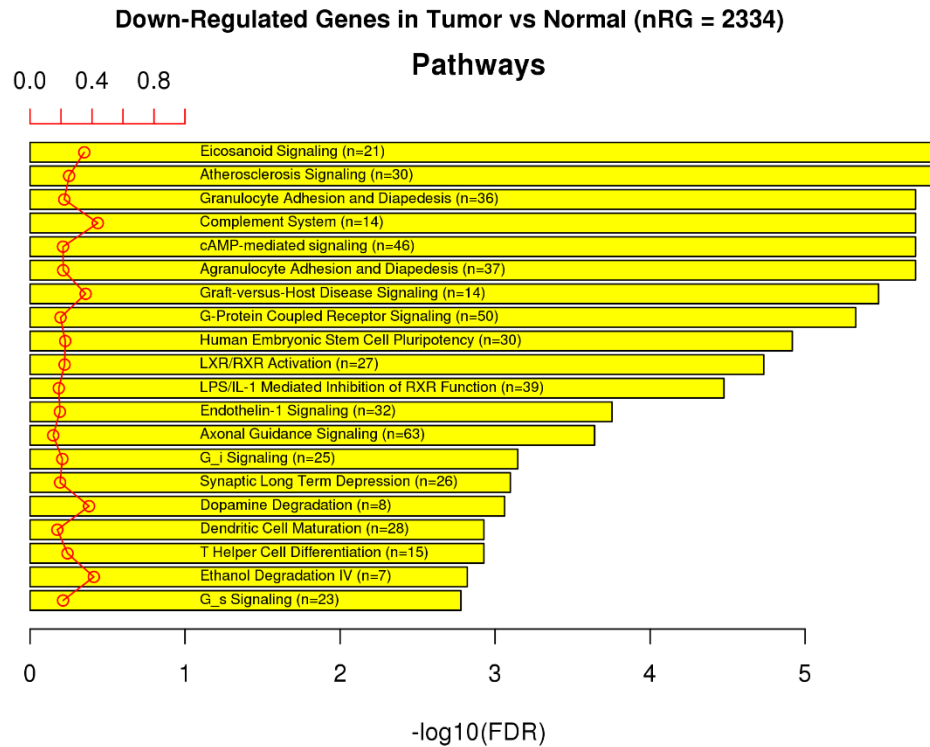


Figure 3.19. Downregulated pathways of gene expression in tumor samples of 55 LUAD patients

3.3.2. DEA of 510 LUAD patients

The differentially expressed genes (DEGs) in tumor samples of 510 LUAD patients are showed in volcano plot in Figure 3.20, as dots of which red ones represent upregulated and green ones downregulated more than two-fold ($\log_2=1$) significantly (FDR corrected p-value < 0.01). As a result of this analysis, 3924 genes are dysregulated more than two-fold with 0.01 FDR significance.



Figure 3.20. Volcano plot of differential expressed genes in tumor samples of 510 LUAD patients

Table 3.7 and Table 3.8 shows the top 10 down-regulated and up-regulated genes which are differentially expressed genes (DEGs) determined by limma and edgeR R packages with their log Fold Change (logFC), adjusted p-value (q-value), Entrez gene IDs and HGNC symbols after enrichment analysis. The list of DEGs was used for pathway analysis and active subnetwork analysis later.

Table 3.7. Top ten significant down-regulated genes in tumor samples of 510 LUAD patients

ensembl_gene_id	entrezgene	hgnc_symbol	logFC	adj.P.Val
ENSG00000108576	6532	SLC6A4	-7.67469368418657	3.11183969281174e-166
ENSG00000170323	2167	FABP4	-6.13109509420139	1.47577195012241e-150
ENSG00000150625	2823	GPM6A	-5.90630829827409	3.78997826923491e-142
ENSG00000135604	8676	STX11	-3.23720948061696	1.95289290988756e-139
ENSG00000182010	219790	RTKN2	-4.56205251096279	1.08582522305812e-136
ENSG00000163815	7123	CLEC3B	-4.31528559475131	3.1615234145803e-130
ENSG00000159307	80274	SCUBE1	-4.33215555670602	3.07001038778257e-129
ENSG00000120156	7010	TEK	-3.44083637544828	8.56115271697181e-129
ENSG00000243566	105375355	UPK3B	-5.80479974416401	9.58417317176578e-129
ENSG00000168309	11170	FAM107A	-4.93054396691964	4.1392246937665e-128

Table 3.8. Top ten significant up-regulated genes in tumor samples of 510 LUAD patients

ensembl_gene_id	entrezgene	hgnc_symbol	logFC	adj.P.Val
ENSG00000135052	51280	GOLM1	2.46618126007993	4.38508086686107e-68
ENSG00000059573	5832	ALDH18A1	1.6002053414484	2.72665457873948e-65
ENSG00000183010	5831	PYCR1	3.57847160020937	1.66280169366614e-63
ENSG00000155660	9601	PDIA4	1.67575262436896	1.94002354761119e-60
ENSG00000162062	80178	TEDC2	3.25696536004962	6.11443772021359e-60
ENSG00000248243	NA	LINC02014	3.52848457496359	8.87153117829031e-60
ENSG00000204949	NA	FAM83A-AS1	5.34724542673934	1.52595746392679e-59
ENSG00000096063	6732	SRPK1	1.53610607737565	1.72932400960279e-56
ENSG00000262772	NA	LINC01977	4.09830628290609	3.37092638295433e-55
ENSG00000128050	10606	PAICS	1.72083035382469	7.055299759656e-53

The most significant dysregulated pathways of DEGs in tumor samples of 510 LUAD patients are including Inhibition of Matrix Metalloproteases, EIF2 Signaling, cAMP-mediated Signaling, Colorectal cancer metastasis signaling, and Human embryonic stem cell pluripotency, which are related to cancer (Figure 3.21). But there are also immune system pathways such as Granulocyte/Agranulocyte adhesion and diapedesis, Leukocyte extravasation signaling, Complement system and IL8 Signaling which may be related to cancer mechanisms.

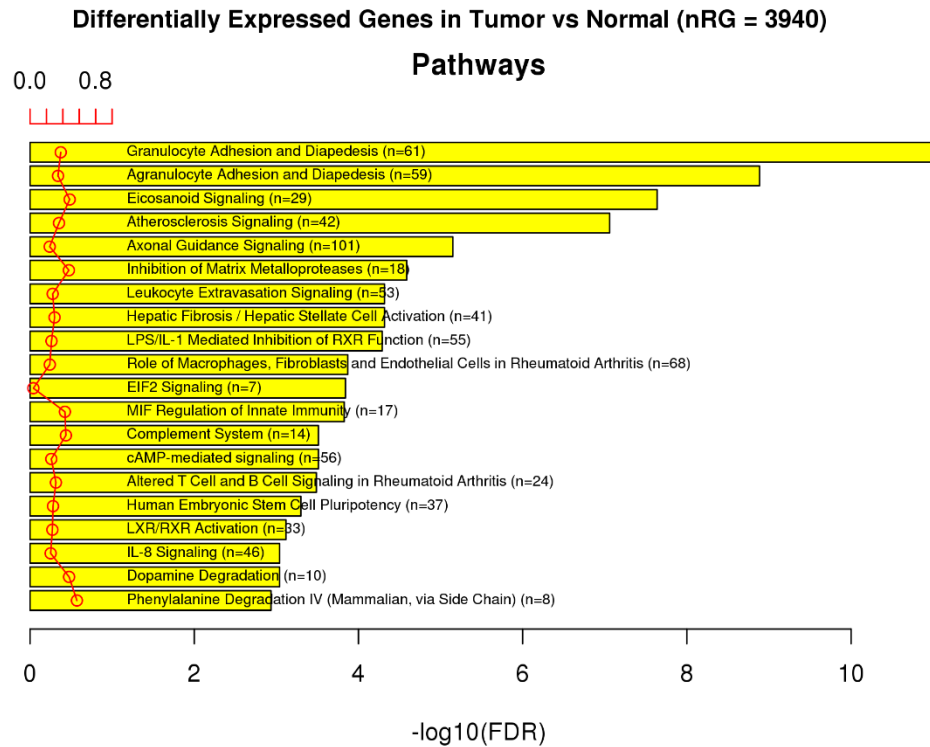


Figure 3.21. Top 20 dysregulated pathways of gene expression in 510 LUAD patients

Most significantly upregulated pathways are Cell Cycle, Mitosis, DNA Damage and Metabolic pathways such as Glycolysis which are expected to be upregulated in tumor cells (Figure 3.22). And the other upregulated pathways such as Hereditary Breast cancer signaling and Bladder cancer signaling supports the impact of upregulated genes in tumorigenesis of LUAD patients.

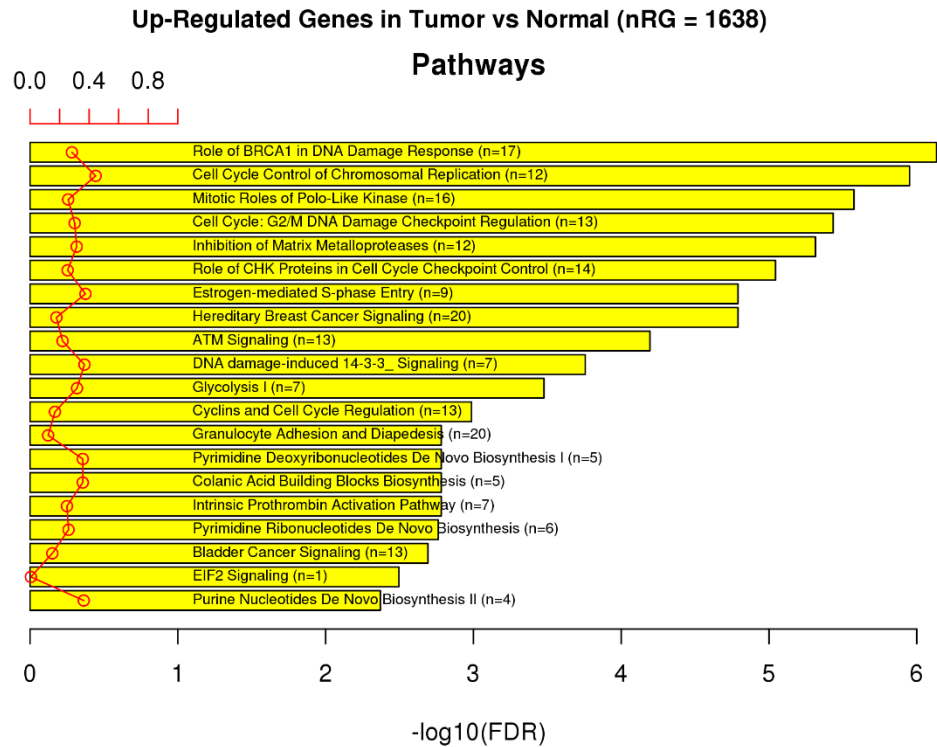


Figure 3.22. Upregulated pathways of gene expression in tumor samples of 510 LUAD patients

The downregulated pathways such as cAMP-mediated signaling, G-protein coupled receptor signaling, VEGF Family Ligand-Receptor Interactions, Human Embryonic Stem Cell Pluripotency and Protein Kinase A Signaling, may be directly related to tumorigenesis (Figure 3.23). cAMP signaling is one of the important mediators playing roles in metabolism, differentiation, proliferation, apoptosis while G-protein coupled receptor signaling, VEGF Family, Human Embryonic Stem Cell Pluripotency and Protein Kinase A Signaling are highly related to cancer. The downregulation of other pathways may be directly or indirectly related to tumorigenesis, tumor growth or other mechanisms resulting in hallmarks of cancer.

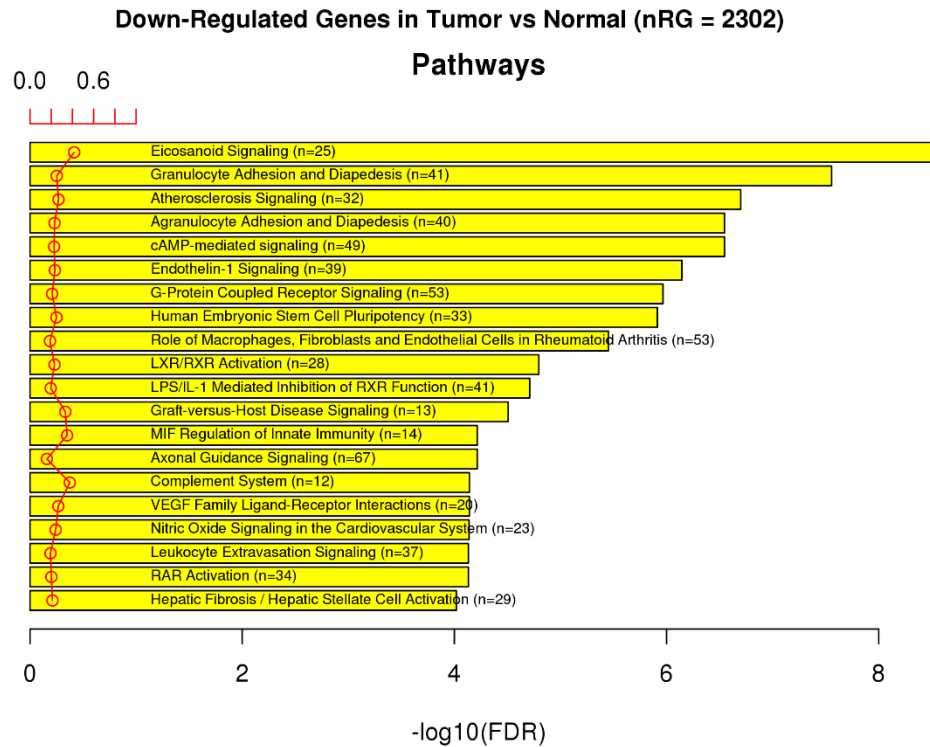


Figure 3.23. Downregulated pathways of gene expression in tumor samples of 510 LUAD patients

3.4. Active Subnetwork and Pathway Analysis

The output of Differentially Expression Analysis (DEA) containing differentially expressed genes with their Ensembl IDs and adjusted p-values (q-values) were used as input of DEsubs package. The active subnetworks of differentially expressed genes in tumor samples of both 55 and 510 LUAD patients were determined by DEsubs package and results were represented as graphs at subnetwork and organism levels.

3.4.1. Active subnetwork and pathway analysis of 55 LUAD patients

In Figure 3.24, the top ten significant genes which play a role in determined subnetworks are represented with their Q-values. These genes are FABP4, WNT3A, EDNRB, TEK, AGER, EPAS1, ACADL, PDIA4, ANGPT4, KL. And from these genes, EPAS1 is transcription factor which has high TF score; KL, EDNRB, FABP4,

AGER, and EPAS1 are related to diseases; EDNRB and TEK proteins are related to drugs in DrugBank and six of them are represented in KEGG pathways (Figure 3.25).

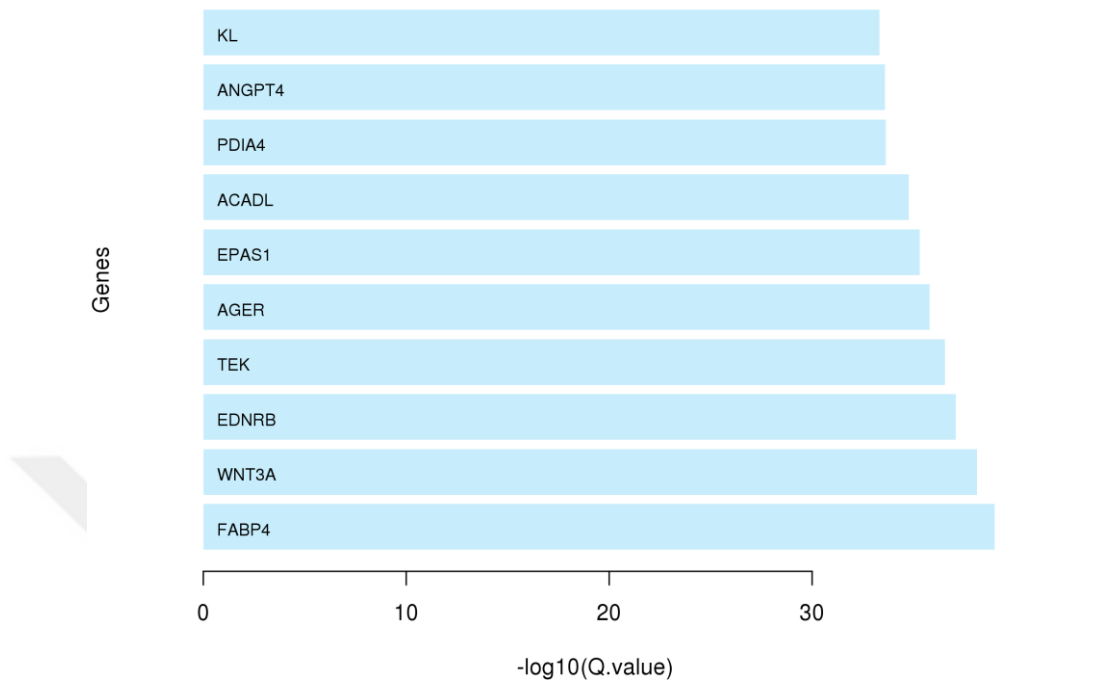


Figure 3.24. Top 10 significant subnetwork genes in tumor samples of 55 LUAD patients

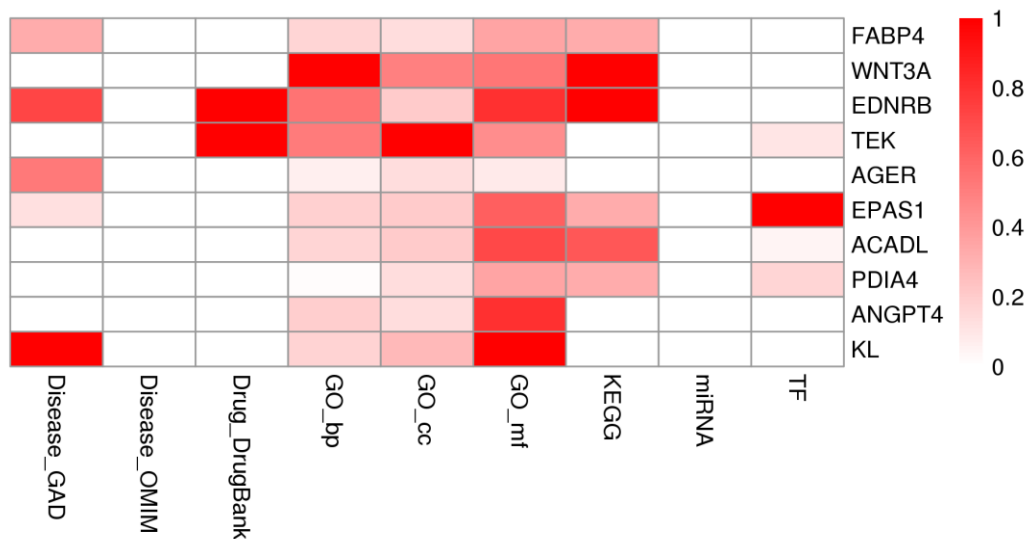


Figure 3.25. Heat map showing functional measurements of top 10 subnetwork genes in tumor samples of 55 LUAD patients

In this analysis, 35 subnetworks were determined and the first three subnetworks are presented in Figure 3.26, 3.27 and 3.28.

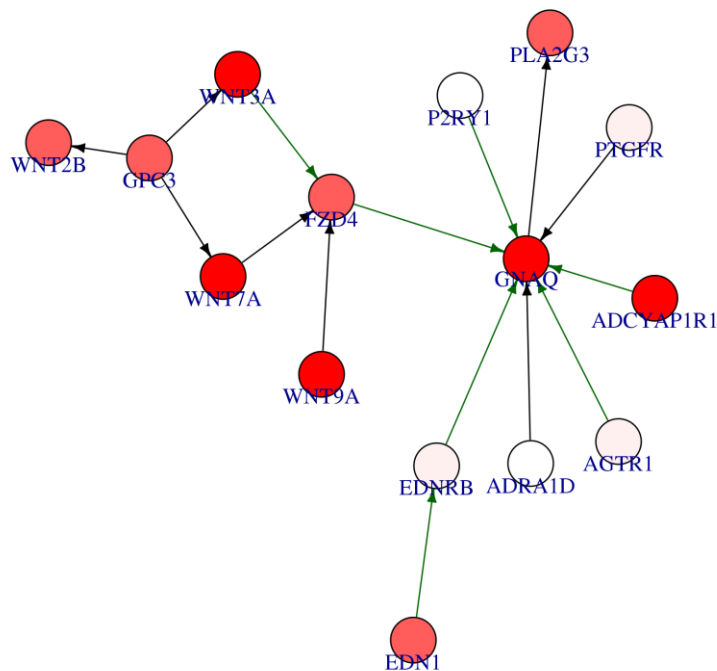


Figure 3.26. Graph of subnetwork 1 from 55 LUAD patients

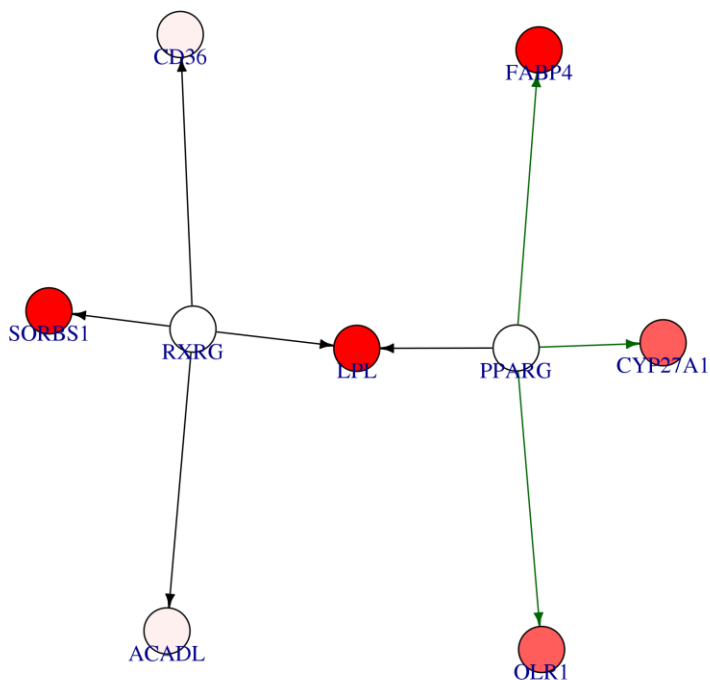


Figure 3.27. Graph of subnetwork 2 from 55 LUAD patients

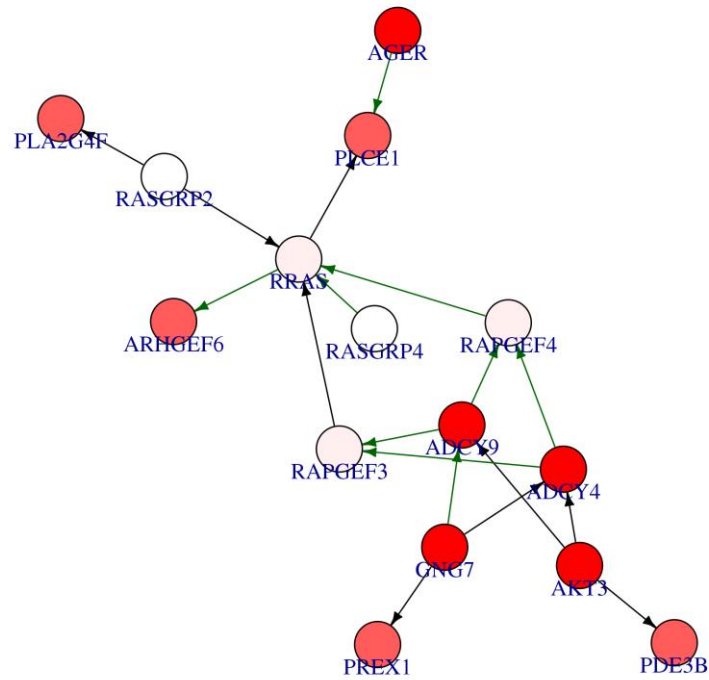


Figure 3.28. Graph of subnetwork 3 from 55 LUAD patients

When we look at the subnetworks' graphs, the prominent genes are WNT genes in subnetwork 1 (Figure 3.26) and AKT3 gene in subnetwork 3 (Figure 3.28). WNT pathway is a major evolutionary conserved signaling pathway playing role in cell differentiation, cell migration, and organogenesis during development and highly related to lung cancer. And AKT3 is one of the AKT family members which play a role in tumorigenesis and are modulators of several tumors.

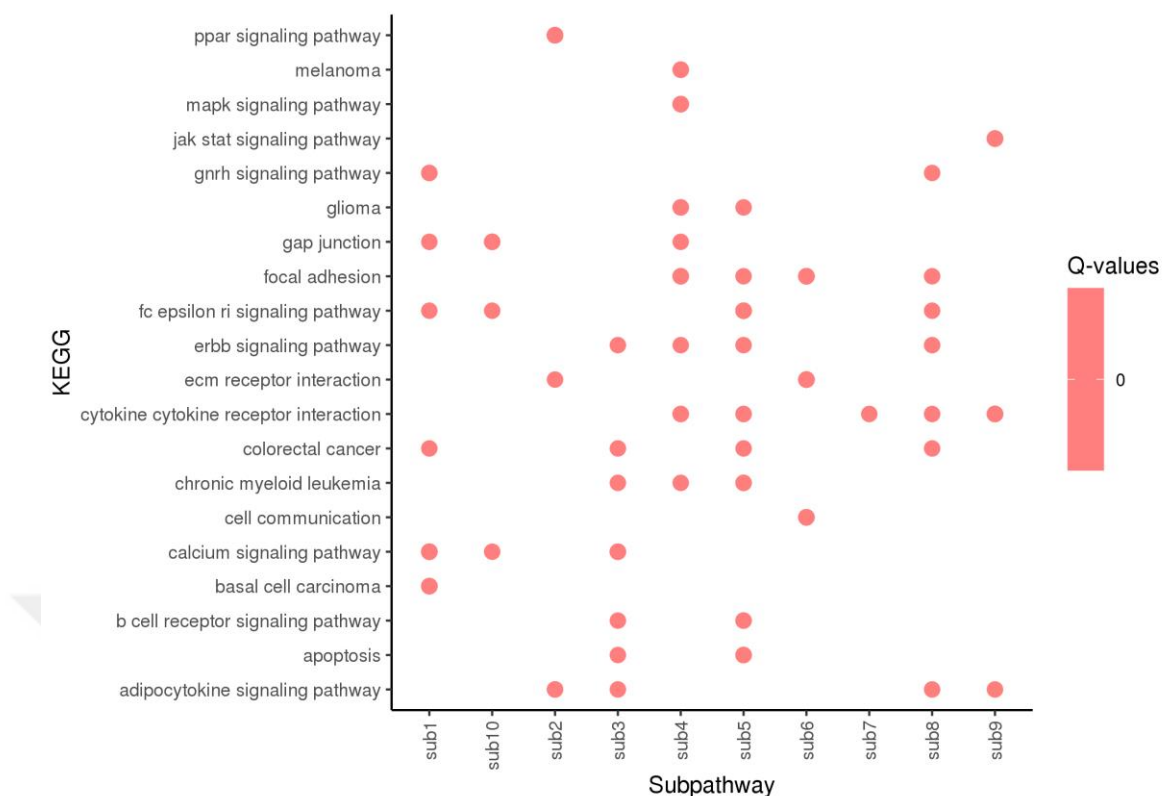


Figure 3.29. KEGG pathways which include subnetwork genes of 55 LUAD patients

The pathways of subnetwork genes are mostly cancer-related pathways such as melanoma, glioma, colorectal cancer, chronic myeloid leukemia, basal cell carcinoma, apoptosis, erbb signaling, jak stat signaling, and map kinase signaling pathways (Figure 3.29). The pathway analysis of subnetwork genes seems to give more accurate pathways which are the strongest candidates in relationship with cancer and as targets in cancer therapy. The transcription factors which are related to subnetworks in Figure 3.30 are also determined and STAT genes, MYC, HIF1A, ELK genes, and BRCA1 are highly related to cancer processes. The drugs which are related to subnetwork proteins in Figure 3.31, may be potential candidates for lung adenocarcinoma therapy.

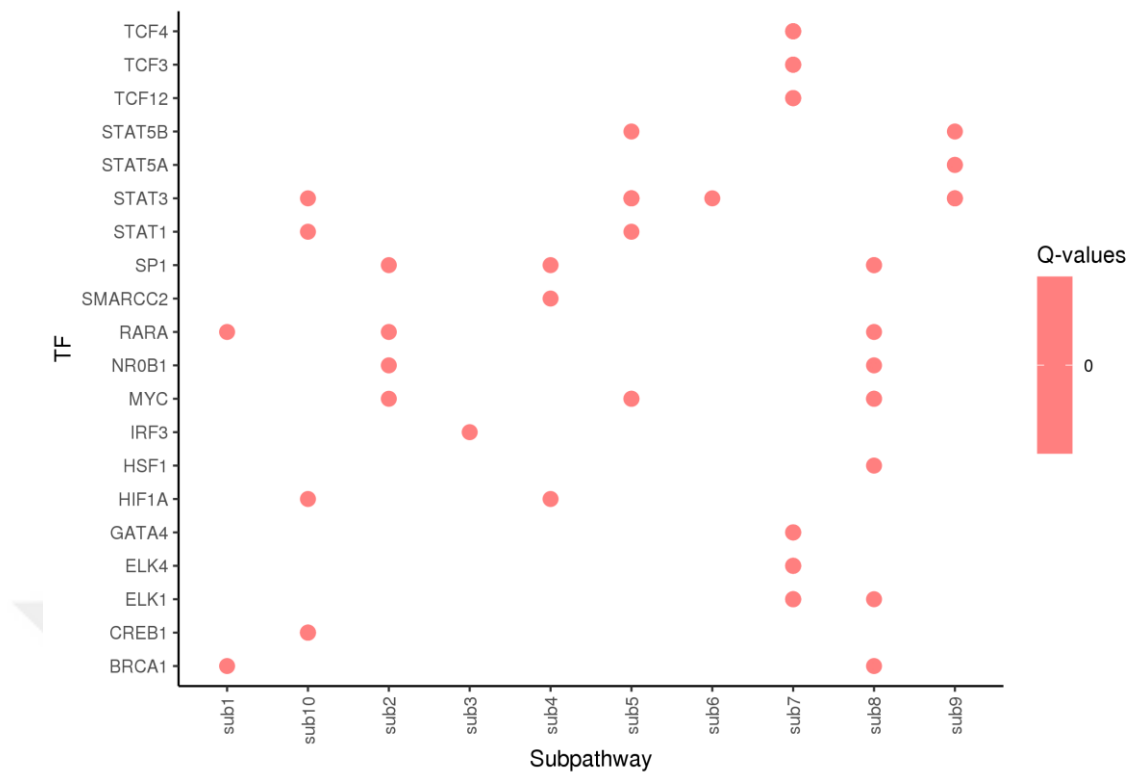


Figure 3.30. TFs which are involved or related to subnetworks of 55 LUAD patients

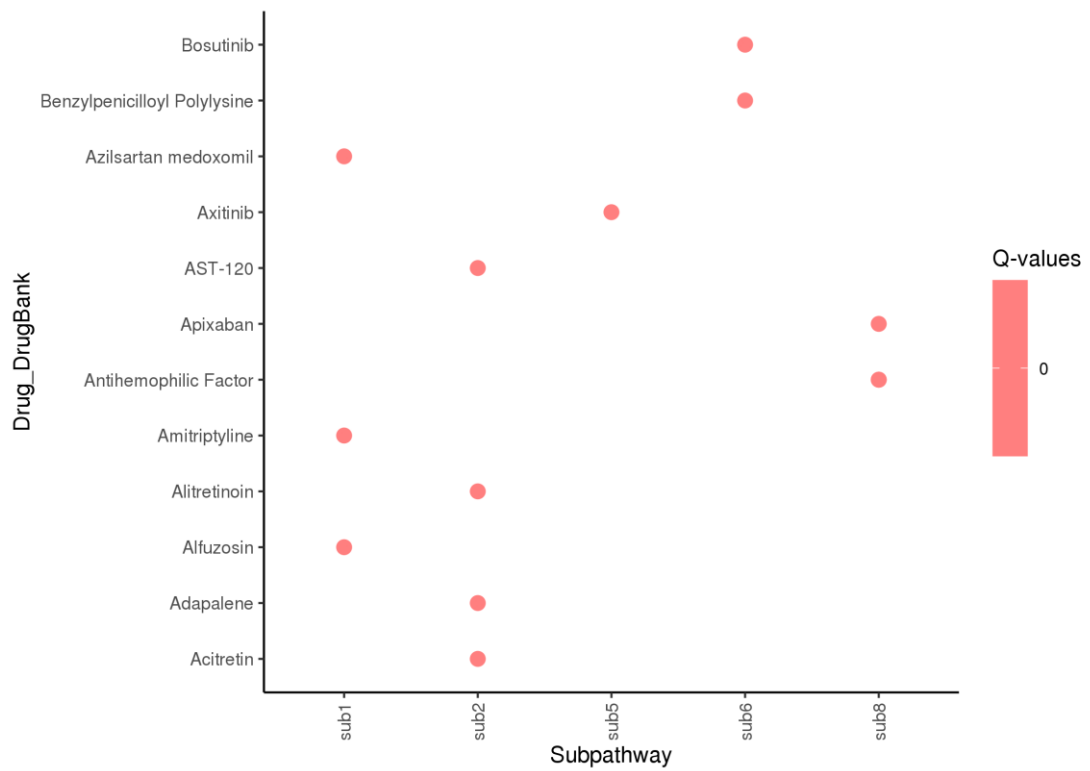


Figure 3.31. Drugs which are related to proteins in subnetworks of 55 LUAD patients

The Entrez gene IDs of the determined subnetworks are extracted from subnetwork results. There are 173 genes which are playing a role in subnetworks identified in this analysis. And these genes will be further analyzed in Cox Proportional Hazards Regression Analysis in order to estimate the effect of them on the survival of the patients.

3.4.2. Active subnetwork and pathway analysis of 510 LUAD patients

The top ten significant genes which play a role in active subnetworks in tumor samples of 510 LUAD patients are represented with their Q-values in Figure 3.32. These genes are TEK, FIGF, EPAS1, IL3RA, COL6A6, SLIT2, ANGPT1, FGR, TNXB and FGFR4. And from these genes, EPAS1 is transcription factor which has high TF score; TNXB and FGFR4 are related to diseases; TEK, FGFR4, and IL3RA proteins are related to drugs in DrugBank and seven of them are represented in KEGG pathways (Figure 3.33).

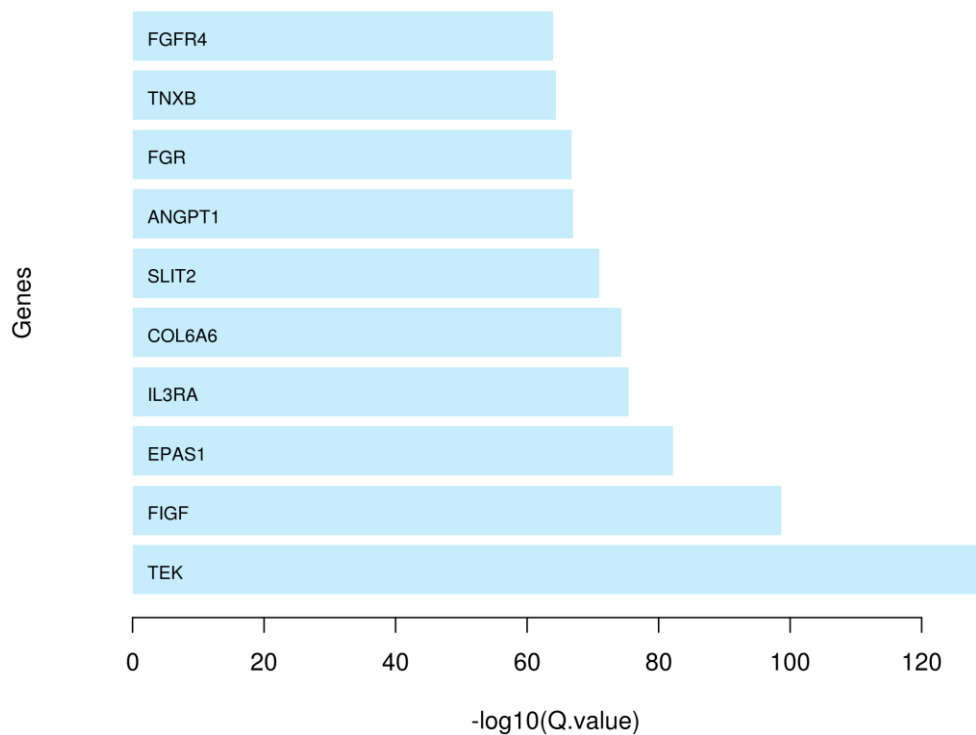


Figure 3.32. Top 10 significant subnetwork genes in tumor samples of 510 LUAD patients

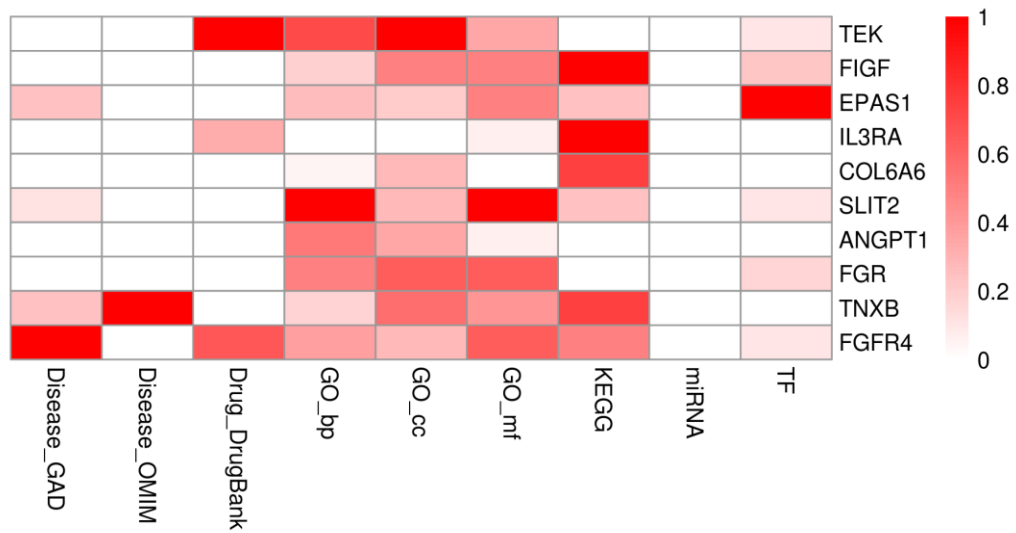


Figure 3.33. Heat map showing functional measurements of top 10 subnetwork genes in tumor samples of 510 LUAD patients

In tumor samples of 510 LUAD patients, only 15 subnetworks were determined although 35 subnetworks were determined in tumor samples of 55 LUAD patients and the first three subnetworks are presented in Figure 3.34, 3.35 and 3.36.

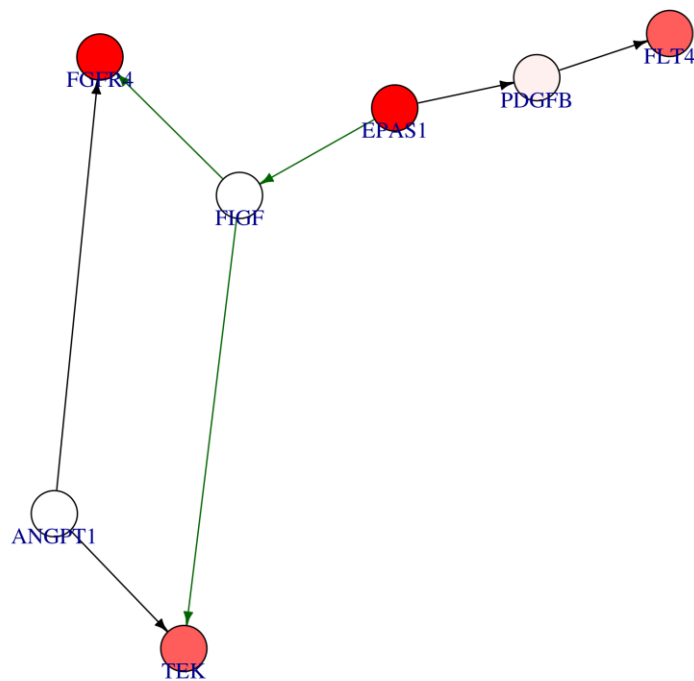


Figure 3.34. Graph of subnetwork 1 from 510 LUAD patients

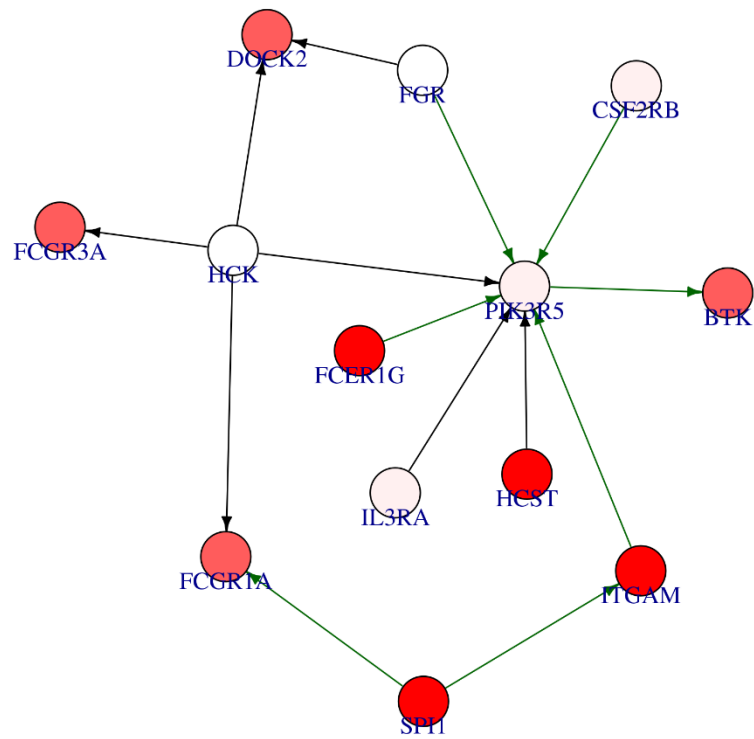


Figure 3.35. Graph of subnetwork 2 from 510 LUAD patients

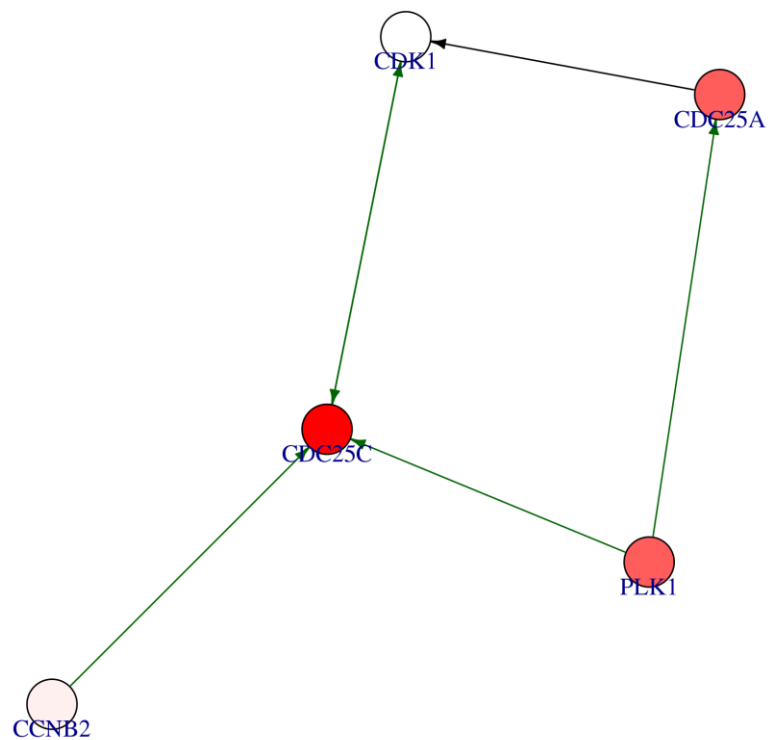


Figure 3.36. Graph of subnetwork 3 from 510 LUAD patients

When we look at the subnetworks' graphs, the remarkable important genes are FGFR4 gene in subnetwork 1 (Figure 3.34), PIK3R5 gene in subnetwork 2 (Figure 3.35) and PIK3R1 gene in subnetwork 3 (Figure 3.36). FGF (Fibroblast Growth Factor) pathway is including highly conserved growth factors and receptors playing role in cell proliferation, cell migration, survival, angiogenesis, and organogenesis through MAPK and PI3K pathways and highly related to lung cancer. And PIK3R genes are regulatory subunits of PI3K complex which is involved in tumorigenesis and other cancer mechanisms.

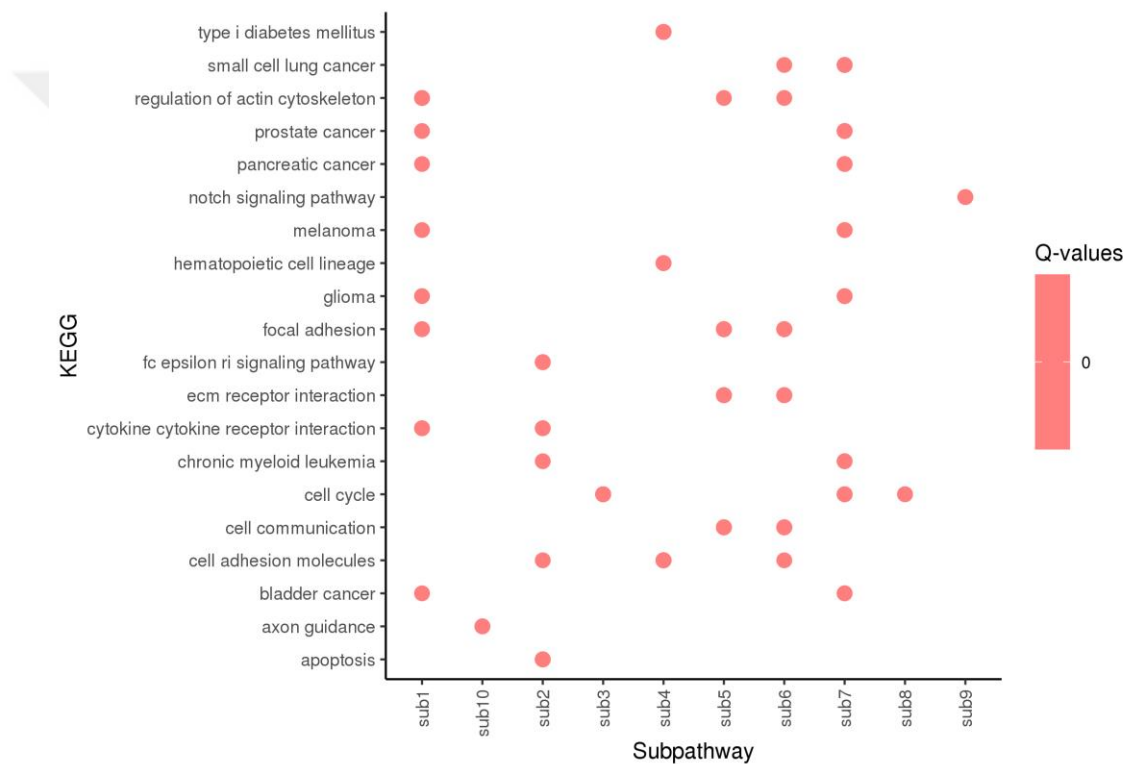


Figure 3.37. KEGG pathways which include subnetwork genes of 510 LUAD patients

The pathways of subnetwork genes are mostly cancer-related pathways such as small cell lung cancer, prostate cancer, pancreatic cancer, melanoma, glioma, erbb signaling pathway, chronic myeloid leukemia, cell cycle, bladder cancer and apoptosis (Figure 3.37). The other pathways such as regulation of actin cytoskeleton, focal adhesion, ECM receptor interaction, cell adhesion molecules, cell communication, cytokine-cytokine receptor interaction, and b cell receptor signaling pathway are most probably

related with metastasis and immune system regulation of cancer. The transcription factors which are related to subnetworks in Figure 3.38 are also determined and STAT and SMAD genes, MYC, CREB1 genes, and BRCA1 are highly related to cancer processes. The drugs which are related to subnetwork proteins of two subnetworks may be potential candidates for lung adenocarcinoma therapy (Figure 3.39).

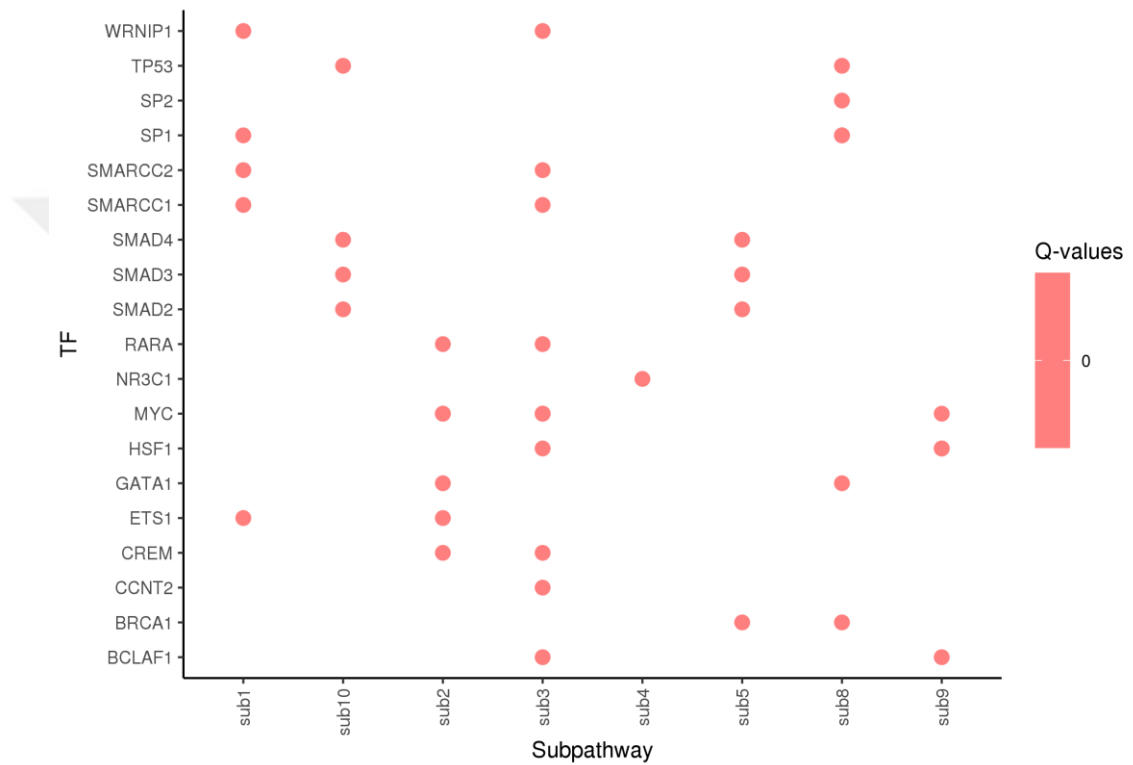


Figure 3.38. TFs which are involved or related to subnetworks of 510 LUAD patients

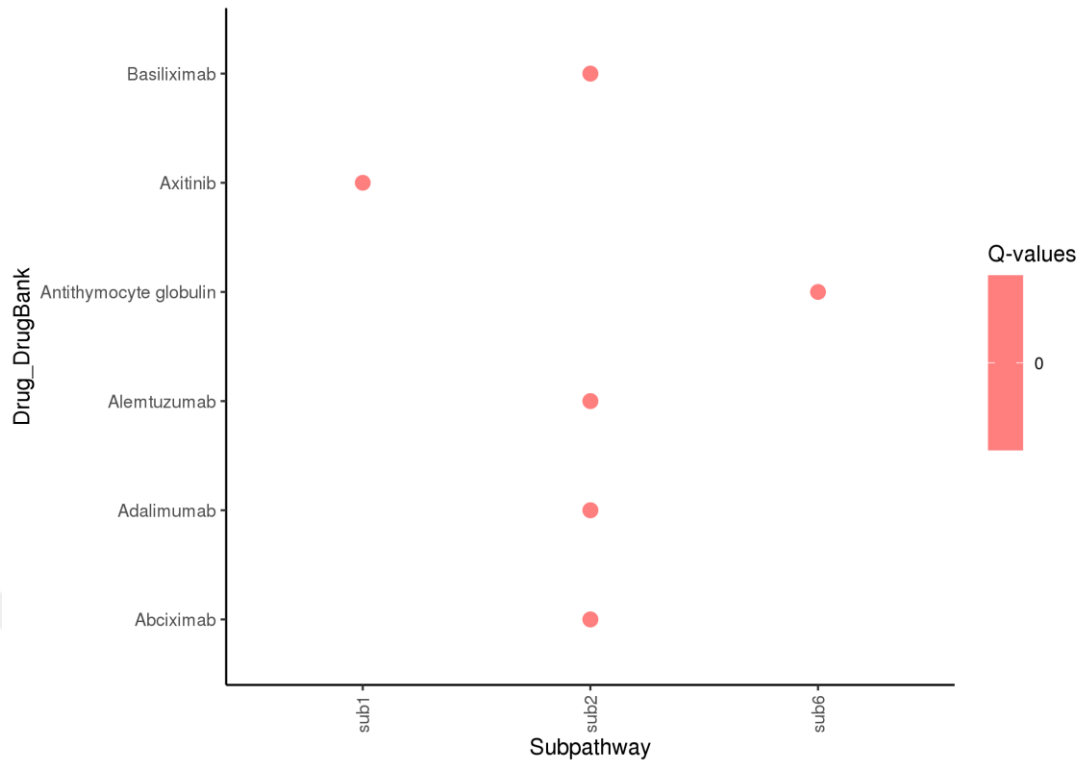


Figure 3.39. Drugs which are related to proteins in subnetworks of 510 LUAD patients

The Entrez gene IDs of the determined subnetworks were extracted from subnetwork results. There are 66 genes which are playing a role in subnetworks identified in this analysis. And these genes will be further analyzed in Cox Proportional Hazards Regression Analysis in order to estimate the effect of them on the survival of the LUAD patients.

3.5. Statistical Analysis

Clinical data of patients were downloaded from the TCGA database using the *TCGAbiolinks* package and the data of patients of interest was extracted. Survival time after diagnosis and vital status with major clinical features such as gender, age, tumor stage and years of smoking from the clinical patient records were extracted for further analysis. Univariate Cox Proportional Hazards Regression (CPHR) analysis was performed for clinical features, significant SNVs, CNVs, and DEGs in active

subnetworks. Then, multivariate Cox Proportional Hazards Regression (CPHR) analysis was performed for significant variates from univariate CPHR analysis. Patients were clustered based on the expression of DEGs in active subnetworks, which are significant results of multivariate CPHR analysis. Kaplan–Meier survival plot was generated for these clusters and the statistical significance of the results was determined using the log-rank test.

3.5.1. Statistical analysis of 55 LUAD patients

Firstly, univariate Cox Proportional Hazards Regression (CPHR) was performed for the clinical features such as gender, age, tumor stage and years of smoking of 55 LUAD patients, summarized in Table 3.9. None of these features can affect survival time significantly. Tumor stage 2 and tumor stage 3 versus tumor stage 1 were significant but the test of univariate CPHR was not significant for this feature. Therefore, these clinical features were not used for multivariate CPHR analysis.

Table 3.9. Summary of clinical features of 55 LUAD patients

Category	Number
Age at diagnosis (median; range)	66 (42-86)
Gender	
Female	33
Male	22
Tumor stage	
I	28
II	12
III	12
IV	2
NA	1
Vital status	
Alive	31
Dead	24

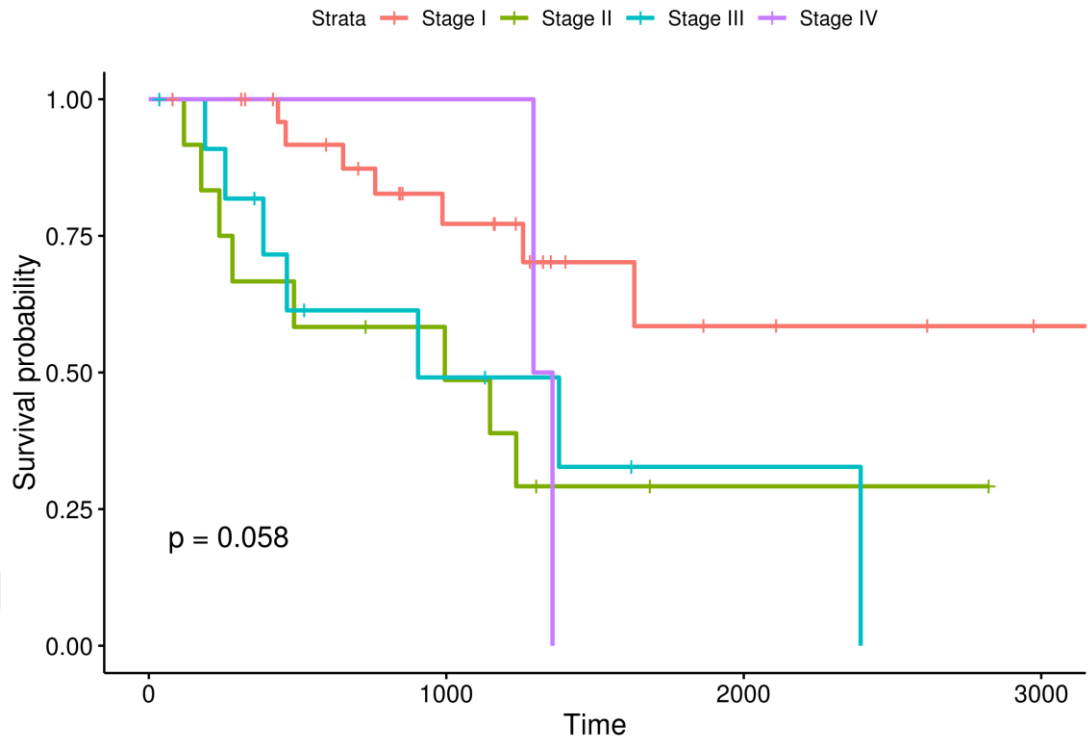


Figure 3.40. Overall survival plot for tumor stages of tumor samples of 55 LUAD patients

Univariate CPHR analysis was performed for significant SNVs, CNVs and active subnetwork DEGs identified at previous steps. Then multivariate CPHR was performed for significant SNVs, CNVs and subnetwork DEGs from univariate CPHR. The candidate driver genes identified by multivariate CPHR includes SNVs (four of six significant SNVs) and active subnetwork DEGs but not any CNVs. RB1, EGFR, STK11 and KRAS mutations were determined to be important with expressions of CD74, ANGPT1, PTGER4, NR3C2, RAPGEF3, PLA2G3, GPC3, WNT3A, GNG7, VEGFD, COL4A3, C2, KLRD1, ARHGEF6, COL6A5, KLF2, LPL, LEPR, CACNA1D, IL11RA, PRKCE, SCN4B, ROBO2, PDE3B and CD244 genes (Table 3.10).

Table 3.10. Candidate driver genes determined by multivariate CPHR, of 55 LUAD patients

Gene Symbol	HR**	p value	Gene Symbol	HR**	p value
CD74	6.97E-135	4.393E-13	KLRD1	1.19E-44	0.0019095
ANGPT1	3.71E-118	7.273E-12	ARHGEF6	8.05E+048	0.0036777
PTGER4	3.55E+116	5.047E-11	COL6A5	2.96E+046	0.0039327
NR3C2	2.29E-105	3.122E-10	KLF2	3.27E+066	0.0042510
RAPGEF3	2.38E-131	1.024E-09	LPL	1.26E-48	0.0082011
PLA2G3	1.04E+056	7.001E-09	LEPR	9.20E+047	0.0109286
GPC3	2.14E+054	1.006E-05	KRAS**	7.84E+161	0.0134146
WNT3A	6.94E-54	2.554E-05	CACNA1D	3.71E-52	0.0140458
RB1**	1.61E-284	0.0001716	IL11RA	4.60E+051	0.0147390
GNG7	6.61E-78	0.0002045	PRKCE	7.33E-77	0.0162075
EGFR**	5.51E-163	0.0002331	SCN4B	3.89E+043	0.0291289
STK11**	2.01E-238	0.0003000	ROBO2	1.32E+033	0.0344102
VEGFD	1.92E+034	0.0003099	PDE3B	9.35E-37	0.0434445
COL4A3	5.63E+041	0.0015219	CD244	1.73E-37	0.0463268
C2	5.65E+059	0.0018156			

* HR: Hazards ratio, ** Mutated Genes identified by SomInaClust

The pathways of significant genes determined by multivariate CPHR includes cancer pathways such as Glioma Signaling, Glioblastoma Multiforme Signaling, Colorectal Cancer Metastasis Signaling, Ovarian Cancer Signaling, Molecular Mechanisms of Cancer, Non-small Cell Lung Cancer Signaling, Breast Cancer Regulation by Stathmin1, HER2 Signaling in Breast Cancer and highly lung cancer related pathways such as G Beta Gamma Signaling, Phospholipase C Signaling, IL8 Signaling, Thrombin Signaling, G-Protein Coupled Receptor Signaling, UVC-induced MAPK Signaling and ERK/MAPK Signaling (Figure 3.41).

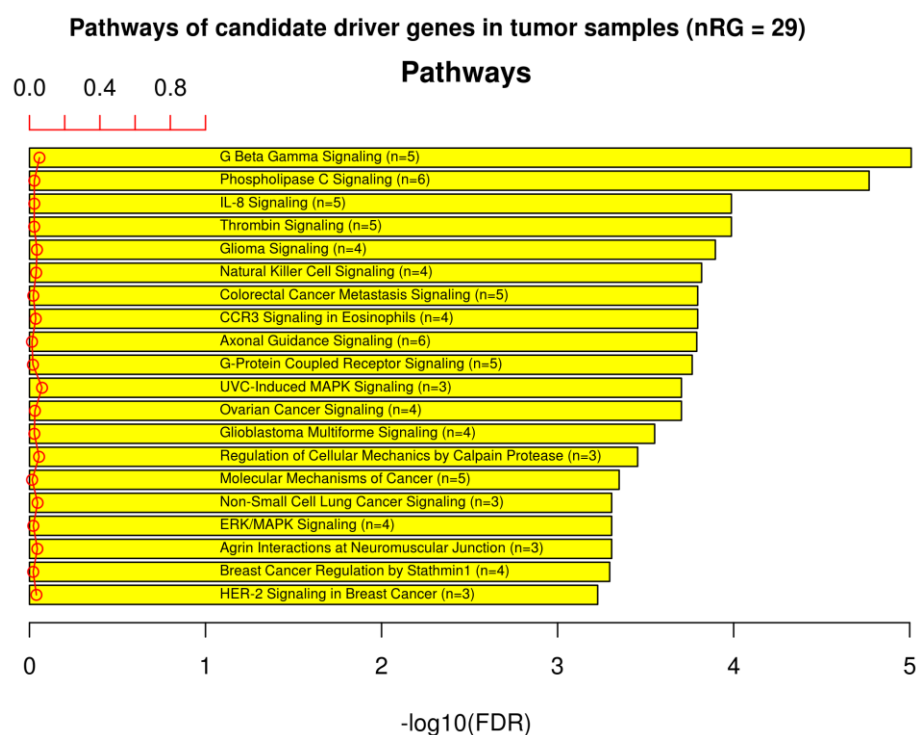


Figure 3.41. Pathways of significant genes determined by multivariate CPHR, of 55 LUAD patients

Heatmap of expression levels of significant subnetwork DEGs determined by multivariate CPHR analysis (Table 3.10), represents apparent dichotomous pattern (Figure 3.42). The patients were clustered into two by hierarchical clustering and survival analysis was performed for these clusters of patients. Survival plot shows that there is a significant divergence between the survival probabilities of two clusters of patients (Figure 3.43). The patients at the upper panel of heatmap showed poorer prognosis significantly in survival plot.

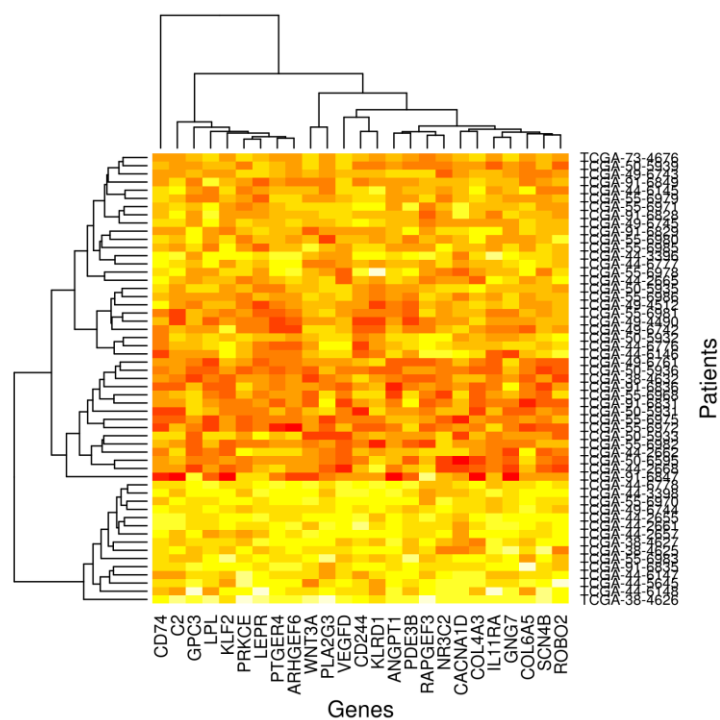


Figure 3.42. Heatmap of significant DEGs from multivariate CPHR for 55 LUAD patients

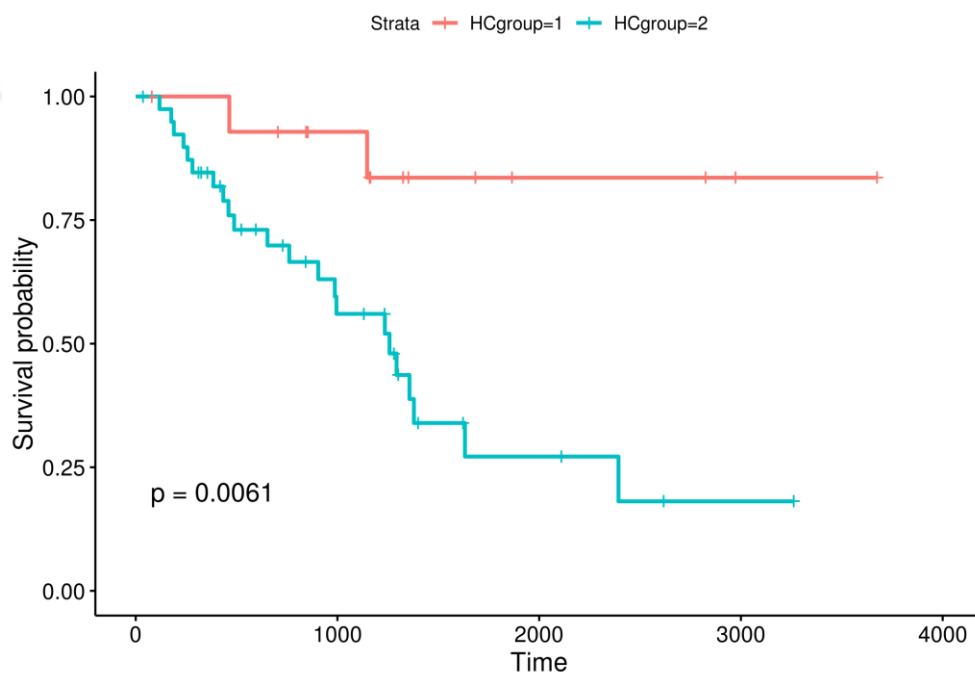


Figure 3.43. Overall survival plot of hierarchical clusters of 55 LUAD patients

3.5.2. Statistical analysis of 510 LUAD patients

Univariate Cox Proportional Hazards Regression (CPHR) was performed for the clinical features such as gender, age, tumor stage and years of smoking of 510 LUAD patients (Table 3.11). Only tumor stage can affect survival time significantly (Figure 3.44). Therefore, tumor stage was used with other variates for multivariate CPHR analysis.

Table 3.11. Summary of clinical features of 510 LUAD patients

Category	Number
Age at diagnosis (median; range)	66 (33-88)
Gender	
Female	273
Male	237
Tumor stage	
I	275
II	119
III	84
IV	25
NA	7
Vital status	
Alive	326
Dead	184

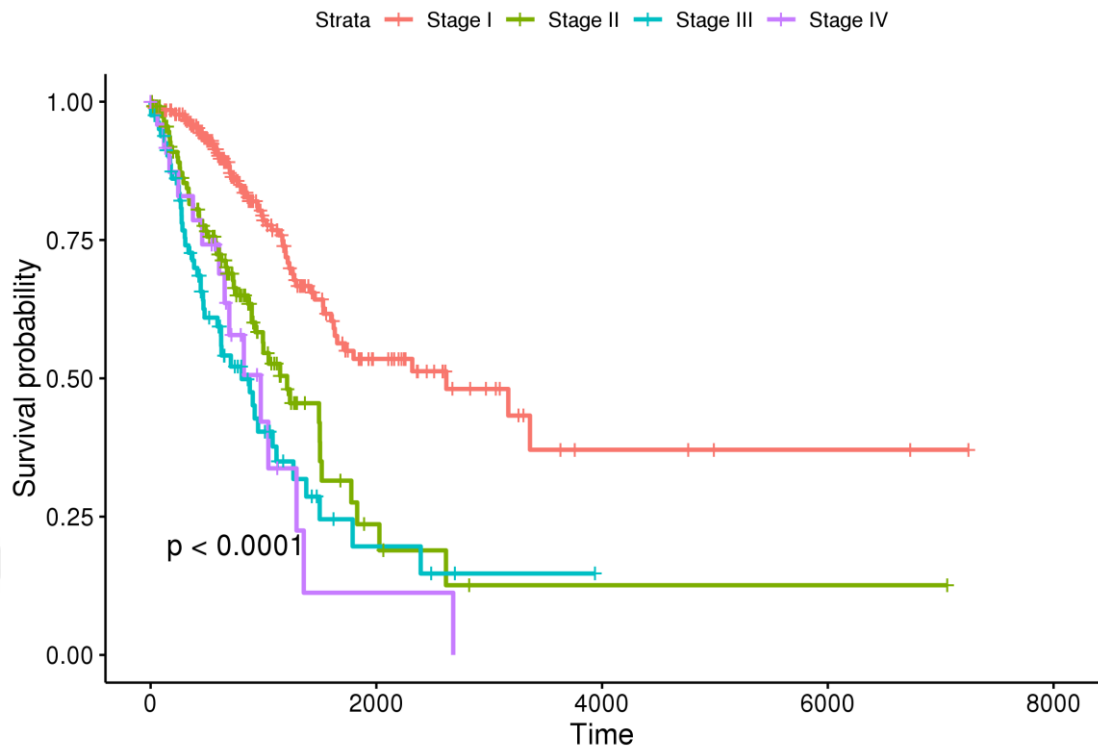


Figure 3.44. Overall survival plot for tumor stages of tumor samples of 510 LUAD patients

Univariate CPHR analysis was also performed for significant SNVs, CNVs and active subnetwork DEGs of 510 LUAD patients, which were identified at previous steps. Then multivariate CPHR was performed for significant SNVs, CNVs and subnetwork DEGs from univariate CPHR. The candidate driver genes and clinical feature identified by multivariate CPHR includes 303 CNVs, 41 SNVs, 8 active subnetwork DEGs and tumor stage as a clinical variable (Table 3.12, Table 3.13 and Table 3.14).

Table 3.12. Impactful CNVs determined by multivariate CPHR, of 510 LUAD patients

Gene Symbol	Aberration	HR*	p value	Gene Symbol	Aberration	HR*	p value
ELDR	Amp	321734.892412471	0	TKTL1	Amp	14523188263.8246	0
AC095038.2	Amp	3.7125106019909E-151	0	AC069218.2	Amp	693370695.749558	0
AC091770.1	Amp	9.8283682894887E-05	0	RAB3IP	Amp	43086.1941571561	0
PSPH	Amp	22315.6182042742	0	AP000457.1	Amp	5.41665864055591E-08	0
AC011294.1	Amp	1.18843247740884E+074	0	MME	Amp	75895.6746979764	0
VN1R25P	Amp	1795453.23403042	0	RN7SKP40	Amp	2.25158606865166E-11	0
AC004870.4	Amp	1.04883254292872E-34	0	TTYH3	Amp	725032.4280606	0
AC099654.2	Amp	3.95704051021231E+138	0	RPS6P4	Amp	10222.5376655233	0
TNS3	Amp	1.14519431325776E-77	0	LINC02569	Amp	395572.148223863	0
TRIM60P16	Amp	186576.533915415	0	NEK4P1	Amp	2223539.95549448	0
AC099654.12	Amp	0.000103706920647	0	MME-AS1	Amp	2.51231245731613E-08	0
SPATA48	Amp	1.00240345804185E-06	0	AC012309.1	Amp	3.1952374399005E-05	0
EPS15P1	Amp	1.80818131088951E+154	0	ZNF720	Amp	117690.435929682	0
LINC01447	Amp	462398398102.844	0	RPL21P103	Amp	180898.5999919	0
AC069282.1	Amp	5.10422269724694E+062	0	AL512428.1	Amp	4.54415746872211E-06	0
AC020743.1	Amp	9.38549405702336E-142	0	AC104629.1	Amp	7521.28772756885	0
AC073115.2	Amp	2.06131727726549E-12	0	AC074050.1	Amp	6.36385267757131E-05	0
CCM2	Amp	6.20891518095599E-147	0	AC016590.1	Amp	14332.7364211938	0
RPL7L1P2	Amp	8.67271115864307E+055	0	ABCB5	Amp	0.000841474109268	0

Table 3.12. (continue)

Gene Symbol	Aberration	HR*	p value	Gene Symbol	Aberration	HR*	p value
LINC01445	Amp	1.74453727489554E+030	0	ADCY1	Amp	0.000861962316122	8.55937792321652E-307
HMGN1P19	Amp	6.83061885764903E-200	0	AC020743.2	Amp	1186.86262702605	8.66814551925979E-302
DBNL	Amp	3.20016081203172E+030	0	MICALL2	Amp	887.580574557496	2.75507300611048E-297
IGFBP3	Amp	1.55413306903619E+084	0	HECW1	Amp	0.00073690704981	2.96415018757498E-293
ROBO2P1	Amp	349695971502.375	0	RSPH10B2	Amp	573.464414320499	3.54241659983994E-267
SLC25A5P3	Amp	1.64347782945372E-23	0	NAALADL2-AS3	Amp	0.000181766601658	1.25115657351498E-254
AC004830.1	Amp	2.09874428364667E-82	0	AGR3	Amp	0.002217799898924	2.30112985590526E-246
VSTM2A-OT1	Amp	1.14185985217421E-47	0	GRIFIN	Amp	0.002045918171266	2.85548095215654E-245
GTF2IP13	Amp	4.24480279854668E-49	0	AC020656.1	Amp	1974.45560590514	4.38611590911923E-238
MIR6837	Amp	3.07842871794125E-27	0	MTCO2P33	Amp	6066.14429188264	2.90578889662267E-237
AC011228.1	Amp	4.38061299438559E+060	0	ABCF1	Amp	0.000406061973306	9.08325943107901E-233
AC093775.1	Amp	6.06184542804004E+043	0	SUGCT	Amp	0.002107063752831	4.539807287572E-218
OGDH	Amp	4.35091906024834E+092	0	STEAP1B	Amp	348.240339863623	2.16751435620897E-217
POM121L12	Amp	1.67664550944E+020	0	RPL31P35	Amp	0.005186466412476	7.50237383793533E-186
AC004920.1	Amp	5.30050882881289E+082	0	AC068759.1	Amp	0.001108808142394	1.14337459570248E-184
HAUS6P1	Amp	1.36523678636111E-12	0	AC092848.1	Amp	180.342389064214	2.42933376494324E-182
DDX56	Amp	1.56148576572983E-05	0	AC004899.1	Amp	221.791782359474	6.01079325208481E-181
AC005999.2	Amp	1.54513900183933E-05	0	AC020637.1	Amp	0.000597583299545	4.26118936829474E-178
AC079763.1	Amp	1.36442006240131E-05	0	AC004949.1	Amp	0.006245756679289	1.14299802762219E-164
STK17A	Amp	5.08094368556832E-110	0	AC002480.1	Amp	0.00729438157834	7.20328747897469E-161

Table 3.12. (continue)

Gene Symbol	Aberration	HR*	p value	Gene Symbol	Aberration	HR*	p value
UBE2D4	Amp	4.07401744406174E-08	0	AC009500.1	Amp	0.008964979001408	3.17926067167296E-144
RNU2-29P	Amp	1.49619731507511E-21	0	PMS2CL	Amp	112.666981742803	8.52373232071797E-138
IKZF1	Amp	3.88894036618765E+148	0	VOPP1	Amp	0.013194873149354	2.60261099885573E-117
AC073109.1	Amp	8.53206746916539E+147	0	AC092647.5	Amp	0.013423098639128	1.05938980206236E-115
GPNMB	Amp	1.0886284690488E-135	0	RANP1	Amp	0.003661263097168	2.66141579956392E-111
TMEM106B	Amp	3.88748919312066E+116	0	MECOM	Amp	0.005636160891475	2.79455335736291E-110
SNX13	Amp	81455.8022236487	0	COBL	Amp	0.014972143936538	7.08990948448854E-107
AC004160.1	Amp	3.29236210978842E+052	0	AC006150.1	Amp	49.8115893787579	3.84264931154477E-102
AC011287.2	Amp	5.49642437265859E-70	0	PKD1L1	Amp	48.3891221201434	1.15657759978334E-94
AHR	Amp	2.87015298037405E-106	0	VPS8	Amp	0.006124992223308	1.24891831072173E-89
RNU1-14P	Amp	4.93410374857768E-12	0	LANCL2	Amp	0.027171145293457	3.91881097908864E-85
AC004415.1	Amp	2.42689852982673E-218	0	RPL7AP25	Amp	0.009018066241776	4.29844311378807E-81
VWDE	Amp	7.49000985840662E+258	0	ARHGAP4	Amp	173.509771095929	7.01916182661868E-77
ETV1	Amp	1.58874521717264E-07	0	AC069431.2	Amp	95.3334616896351	4.33048168819899E-76
ITGB8	Amp	1.69957342656883E+016	0	CAMK2B	Amp	0.031178079312529	8.22281461463549E-76
MALSU1	Amp	1.82611589919231E+128	0	SNORA80D	Amp	0.037628451284902	3.23490102376259E-70
AC004692.1	Amp	4.3514064352286E+170	0	RNU6-1125P	Amp	24.5682448202532	6.65141416868202E-70
TUBG1P	Amp	9.27046000144007E-05	0	AC007009.1	Amp	22.7750541495122	3.53347284198435E-64
AC013470.2	Amp	4.03188015911087E-116	0	AKT2	Amp	0.007450543652906	1.89556400858746E-63
SCIN	Amp	8.98753827342188E+185	0	FAM126A	Amp	0.049401120076133	9.6207445550651E-60

Table 3.12. (continue)

Gene Symbol	Aberration	HR*	p value	Gene Symbol	Aberration	HR*	p value
AC006041.1	Amp	4.76472600659706E+070	0	NAALADL2-AS2	Amp	0.016811557280352	9.02198580042711E-59
AMPH	Amp	19783.5381171428	0	FGFR1OP2	Amp	0.013295939942556	3.31275621157185E-57
SEPT7P2	Amp	3.28725721138467E+056	0	U82695.1	Amp	0.013816621593254	8.58394852980224E-57
NDUFA4	Amp	1.87259673071139E+033	0	GUSBP10	Amp	20.4881453278498	1.53176244196873E-53
AC004160.2	Amp	7.05140285249324E-24	0	SDK1	Amp	16.7858765272571	1.38124583866631E-50
AC013470.4	Amp	4.28757423911973E-298	0	SSPN	Amp	46.968293322821	2.96708083161566E-46
GTF3AP5	Amp	6.73190455639285E-14	0	ITPR2	Amp	0.021042660600784	6.31629391036189E-44
SOSTDC1	Amp	1.73272128870268E+085	0	KRT8P20	Amp	0.080125094815364	2.6459625963144E-40
MACC1	Amp	7.31356907480097E+053	0	AF127577.3	Amp	41.5651421306778	5.04277774260942E-40
AC005682.1	Amp	6.53654744895146E+020	0	RN7SL292P	Amp	11.8220148083178	6.59664680318797E-39
AGMO	Amp	1.25402702764173E-152	0	AL033381.1	Amp	0.057383160841972	9.17985113497464E-38
LINC02442	Amp	7.83099422537304E-15	0	TBC1D30	Amp	27.6923926577442	4.25330315334657E-37
THSD7A	Amp	0.000297284933191	0	AC098592.1	Amp	0.10238892747551	2.64689982736062E-35
MEOX2	Amp	5.592802395265E-12	0	ABCA13	Amp	0.103069536624711	4.25961803414039E-33
MACC1-AS1	Amp	1.00284213070709E-209	0	FRS2	Amp	18.029344951316	1.08830060276028E-32
TOMM7	Amp	154630.850070149	0	MUC22	Amp	16.704517673333	6.68341492872764E-29
AC011891.1	Amp	0.000823139568013	0	PA2G4	Amp	19.0901364310117	1.09878894473503E-28
SFRP4	Amp	3.33514334011458E+020	0	PTPRR	Amp	19.1109647304547	1.42944769452688E-28
DDC	Amp	1.12337576372632E-55	0	AC078841.1	Amp	0.147562547072611	8.2841538482249E-25
RN7SKP266	Amp	4.97894514334581E+094	0	ETFRF1	Amp	12.1384045245437	1.47809555760309E-24

Table 3.12. (continue)

Gene Symbol	Aberration	HR*	p value	Gene Symbol	Aberration	HR*	p value
RPL21P75	Amp	1.51336976655517E-63	0	COA1	Amp	6.94651177962965	2.66927666781016E-24
STARD3NL	Amp	1.19493876313476E-05	0	AC006971.1	Amp	6.03001151360915	3.27072913653006E-23
INHBA	Amp	1.73863094912049E-68	0	IL12A-AS1	Amp	0.08357926244988	6.94816316615838E-22
AC080080.1	Amp	7.78259554788714E-25	0	AL139231.1	Amp	11.1254088325901	7.99424291456405E-22
IGF2BP3	Amp	5.22473777445545E+018	0	RPA3	Amp	5.23508752501737	3.66330971530593E-19
AC005083.1	Amp	3.94366708082657E+104	0	TREX2	Amp	11.4434406096715	2.06502852772015E-18
SNORA20B	Amp	1.10633144629948E+035	0	AC074351.1	Amp	4.51055035142061	1.06038656734855E-16
AC018634.1	Amp	2.26498820216599E-06	0	AC002116.2	Amp	0.123240717084258	3.26391520166235E-16
ZBPB	Amp	55912.3637198408	0	VSTM2A	Amp	0.229037288038992	3.82187124422365E-16
RNMTL1P2	Amp	1.02534113394303E+016	0	AC074397.1	Amp	0.228639768388774	1.53940027557445E-15
HDAC9	Amp	2.51239181859E+068	0	RNU6ATAC42P	Amp	7.09090218298423	1.32453809502615E-14
LINC02408	Amp	2.29943257482131E+017	0	RNU6-279P	Amp	6.60213745271691	3.81337223286572E-14
AC010082.1	Amp	3.61608328662197E+042	0	LEUTX	Amp	8.2462079043572	7.30307818727877E-14
LSM5	Amp	2.69854462130964E-18	0	PSMC1P11	Amp	5.67131536608872	5.27777004629382E-13
MIR1200	Amp	8.917481175121E+019	0	AC007255.1	Amp	0.263045389382844	6.8564551015666E-13
RNU1-15P	Amp	1.61725941229943E-62	0	AC025569.1	Amp	0.176908970032158	3.19670702455692E-12
CCDC126	Amp	1.71709823641965E-20	0	GRID2IP	Amp	3.53540947175866	7.89135964268087E-12
LINC01162	Amp	761635.051144679	0	ZBTB32	Amp	0.213810847701033	4.41988488338329E-10
RAB11AP2	Amp	9.81207538195179E+034	0	AC093520.2	Amp	0.212917986966318	4.83973399795987E-10
BZW2	Amp	2337175.28945645	0	AC017000.1	Amp	2.90043514458262	1.0768759930638E-08

Table 3.12. (continue)

Gene Symbol	Aberration	HR*	p value	Gene Symbol	Aberration	HR*	p value
GRB10	Amp	2.01624299436728E-58	0	AC005294.1	Amp	0.257117872500512	1.91650566391574E-08
JAZF1-AS1	Amp	1.15617427513482E+031	0	AC011287.1	Amp	2.6238541223518	1.44895936435868E-07
CREB5	Amp	2.30555057007611E-19	0	CASC15	Amp	0.272504647553015	4.87348624346551E-06
BMPER	Amp	1.08155960925943E+050	0	MIR5091	Amp	0.262554120277614	8.58719691776066E-05
AC005100.1	Amp	4.74739436468834E-05	0	AC073325.1	Amp	2.07894858149224	9.79741870719975E-05
TAX1BP1	Amp	1.05885801400324E-58	0	RNU4-20P	Amp	2.48055097905346	0.000141253767258
AC005091.1	Amp	7.12025991725107E-24	0	CENPUP2	Amp	2.4825689824917	0.000156991927547
RPS17P13	Amp	7973712.46774032	0	POU6F2-AS2	Amp	1.98882274308384	0.000320498426817
POU6F2-AS1	Amp	1.0465958446524E-35	0	AC022509.2	Amp	2.39934774643442	0.000710228440502
RNU6-1085P	Amp	3.97422893569871E-25	0	LINC01208	Amp	0.456975793021668	0.001711571809568
NTAN1P3	Amp	9.01130466156258E-11	0	FBXL19-AS1	Amp	1.93592796872594	0.008115834012743
JAZF1	Amp	8.88857590283272E+045	0	NXPH1	Amp	0.626898119363789	0.011633317161324
PMS2	Amp	9670.31266743796	0	SPATA13-AS1	Del	7379839861232.15	0
AC090022.1	Amp	701281665510122	0	PHBP13	Del	8886.60720685114	0
TRG-AS1	Amp	66274913.3016981	0	SCAND3P1	Del	2.78580617317553E-10	0
DPY19L1P1	Amp	0.000123160891397	0	LINC00566	Del	1.90018794602663E-09	0
AC073343.2	Amp	226120.534677967	0	AL139081.1	Del	39725.8168732907	0
RF00413	Amp	7.35053305467611E-08	0	COG6	Del	1.58270482832898E-07	0
IKBKG	Amp	2569862.53028536	0	AL160035.1	Del	1.07883857417609E-06	0
AC004982.2	Amp	5.57508474888113E-05	0	LINC01048	Del	550379714.005823	0

Table 3.12. (continue)

Gene Symbol	Aberration	HR*	p value	Gene Symbol	Aberration	HR*	p value
SEPT7	Amp	5381.14193764677	0	AL139327.2	Del	1.60706324362885E+035	0
ELMO1	Amp	0.000358317013664	0	DCLK1	Del	6.46265992171166E-05	0
AP5Z1	Amp	9.34305783586211E-08	0	AL356259.1	Del	8.91901432852004E-35	0
AC087883.1	Amp	2.66066668204137E-16	0	RFC3	Del	420.123257300271	3.22980441178474E-211
AC108721.1	Amp	4.95262745131931E+033	0	FRY	Del	147.831094012522	6.28422714665799E-147
MDM2	Amp	3.45607940078839E+017	0	CCDC169	Del	0.018419617336483	2.61262055200691E-96
BEST3	Amp	4.25126616821217E-06	0	PARP4P2	Del	0.027555860157686	2.56134931027375E-80
AC002511.1	Amp	9.59435507703507E-08	0	RPS3AP44	Del	34.6452815461905	1.92562243951496E-74
ACTB	Amp	4.79466386852619E-15	0	FTH1P7	Del	20.905705990314	9.94213461627386E-57
PDE1C	Amp	3.71005711669223E-11	0	RF02045	Del	10.4911160313052	3.65205524834818E-46
CDK13	Amp	1.06243378436984E-11	0	TPTE2P6	Del	0.082862135629681	1.34722148545008E-37
SLC16A7	Amp	1.05032934741085E-19	0	SLC18A1	Del	9.47802301225356	1.21788051478069E-36
AC025423.3	Amp	1.23225926642577E-16	0	AC084838.1	Del	9.00563798305545	3.79052611615868E-35
BBS9	Amp	27543.5127183098	0	POLR1D	Del	10.6404623850022	2.20242369846321E-33
AC068305.2	Amp	1.96831344373333E-51	0	LINC01058	Del	0.134412490019173	4.53683979804527E-25
FFAR2	Amp	591424911.767088	0	AL157937.1	Del	0.243732089054059	4.83175227808284E-17
G6PD	Amp	3.06094594404453E-18	0	AC090541.1	Del	0.238481650400899	9.15211463101538E-16
EEF1B2P8	Amp	166892584500.26	0	CSGALNACT1	Del	0.258039796072882	3.39948417542306E-14
AC025262.1	Amp	1.08456681106686E+061	0	AC092723.1	Del	0.241391657797528	2.04486090343414E-08
IFNG-AS1	Amp	4.47846824203216E+023	0	UBL3	Del	2.63309445619905	6.41884508050942E-07

Table 3.12. (continue)

Gene Symbol	Aberration	HR*	p value	Gene Symbol	Aberration	HR*	p value
LINC02373	Amp	2.81135829778691E+017	0	LINC02341	Del	0.397134909810499	2.05540340735449E-06
AC025263.1	Amp	3.57211562458684E-21	0	C16orf95	Del	3.31017643077477	3.50111895236332E-06
CYP3A54P	Amp	609376435.789044	0	ADAMTSL1	Del	0.438306260525856	1.49241127761676E-05
MYRFL	Amp	7.55429963581473E-25	0	AC046158.1	Del	0.447516179947664	0.002169263229544
AC068305.1	Amp	1.06725590900479E+048	0	CDKN2B-AS1	Del	1.63122156139024	0.004066450198654
GNS	Amp	6.68595829134991E-25	0	AC068287.1	Del	2.94964081043054	0.015818699648967

* HR: Hazards ratio

Table 3.13. Impactful SNVs determined by multivariate CPHR, of 510 LUAD patients

Gene Symbol	HR*	p value
ATF7IP	347.201470207209	9.17512894764591E-53
LTN1	545.817157886723	2.21325303603351E-38
CDKN1B	0.006450737476465	7.18759799169038E-18
VPS13C	0.07666877304084	2.93403773188146E-16
COL28A1	0.048248661318802	1.26718849443533E-12
IQUB	15.3416847642025	1.12267922185953E-10
DHX15	0.013943545075221	3.09772770142579E-08
FHOD3	6.61967972865553	5.77213955504687E-08
SP1	0.013904411387384	4.10929670910235E-07
WISP3	0.011242833033393	3.55615454893682E-06
NF1	0.323357534921041	5.04913054115691E-06
CTNNB1	0.134011643578054	7.89228478727403E-06
MAP2K4	17.5948581563987	1.31853806181526E-05
LRP1B	1.96068724735922	8.2215452371413E-05
CPVL	0.061916033461516	9.10635922563864E-05
SCAF8	0.140485440292294	9.69159595746002E-05
NFE2L2	5.23965811392576	0.000208493370611
RAD50	13.5490937011725	0.000218823764536
CDKN2A	4.43701695750725	0.000248507015922
TTK	0.163133135939177	0.000426422228571
SPTA1	1.92189477603596	0.000499118287453
KDM5C	0.239774541875573	0.000522981956289
IQGAP2	0.32834483934471	0.0010910703277
KMT2C	2.56537131112732	0.001127929935735
ZNF800	0.23081021606977	0.001428170449342
SMARCA4	2.42130549411532	0.002006452054045
ARID2	2.83084870656274	0.00443188543795
ZBBX	0.354080515461994	0.00472965655629
SENPI	5.63899046554964	0.006622919915981
PNISR	5.81342300585644	0.008512546535538
HMCN1	0.566082787324135	0.01005754455447
DGKB	0.455471771850039	0.012567767100599
B2M	0.190363657974975	0.012587319000111
STK31	2.52525703334771	0.03001307978197
COL11A1	0.6525382095081	0.030129391837473
ARID1A	2.2475040169834	0.032317947666634

Table 3.13. (continue)

Gene Symbol	HR*	p-value
ROCK1	0.45549206045224	0.032601491774405
DST	1.57319683825697	0.041008097599124
BRAF	1.75452449960145	0.04335856776152
NAA15	0.226806319060653	0.045306629517749
KRAS	1.44451387058723	0.047049733167188

* HR: Hazards ratio

Table 3.14. Impactful DEGs and features determined by mv CPHR, of 510 LUAD patients

Gene Symbol / Clinic feature	HR*	p-value
HLA-DOA	0.543229664667665	3.54779991655978E-33
CDC25A	1.69141338447433	5.21568305002388E-18
HLA-DQA1	1.43222438505574	5.68908128477082E-14
BUB1	0.727452357194934	4.58892700315358E-09
ITGA8	1.22132091155711	0.000143989568049
CCNB2	1.16461244432994	0.008326435260666
MAD2L1	1.14966081242955	0.018125897955436
CDK1	0.874213553764107	0.018139093076554
tumor_stage3	11.5622326599801	1.43518390606447E-35
tumor_stage4	3.09951106227159	0.001335208820503

* HR: Hazards ratio

Pathways of significant impactful genes identified by multivariate CPHR are mostly cancer-related pathways such as ATM Signaling, Mitotic Roles of Polo-like Kinase, Cyclins and Cell Cycle Regulation, DNA Damage Signaling, DNA Damage Regulation, Cell Cycle Checkpoint Control, Estrogen Mediated S Phase Entry and GADD45 Signaling. Moreover, immune system pathways such as Graft versus Host Disease Signaling, Antigen Presentation Pathway, T Helper Cell Differentiation, IL-4 Signaling, Altered in T Cell and B Cell Signaling in Rheumatoid Arthritis, iCOS- iCOSL Signaling in T Helper Cells, CD28 Signaling in T Helper Cells, PKC Signaling in T Cells, Dendritic Cell Maturation and Role of NFAT in Regulation of

Immune Response; and also Type I Diabetes Mellitus Signaling are other pathways of impactful genes identified by multivariate CPHR analysis (Figure 3.45).

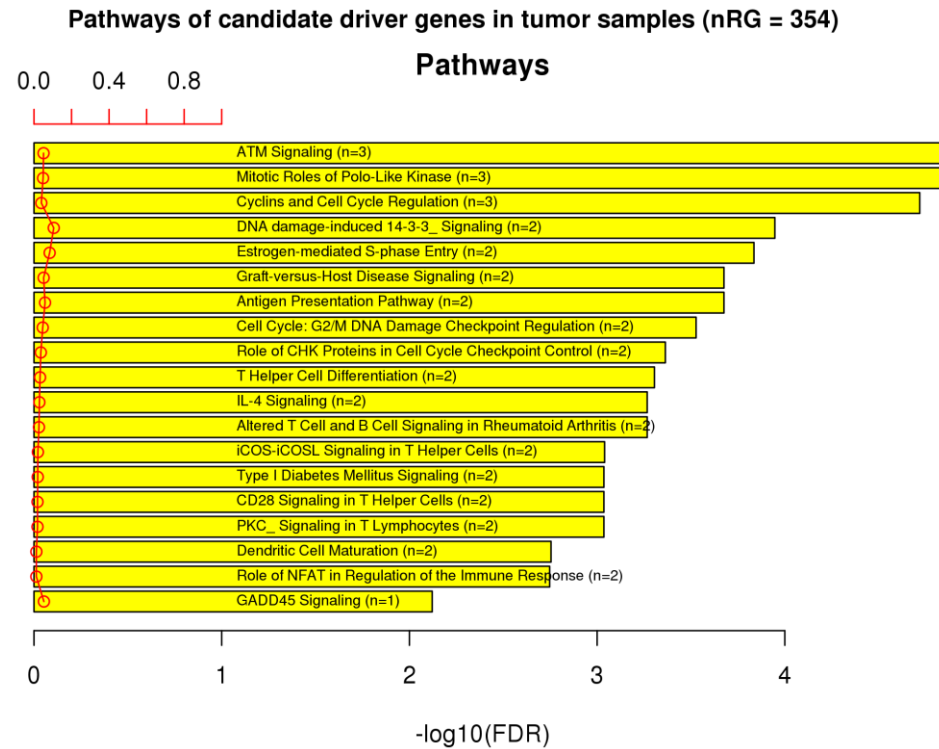


Figure 3.45. Pathways of significant genes determined by multivariate CPHR, of 510 LUAD patients

Heatmap of expression levels of 8 significant subnetwork DEGs determined by multivariate CPHR analysis (Table 3.14), represents apparent dichotomous pattern (Figure 3.46). The patients were clustered into two by hierarchical clustering and survival plot was generated for these clusters of patients. Survival plot shows that there is a similar pattern but significant difference even a little between survival probabilities of two clusters of patients (Figure 3.47).

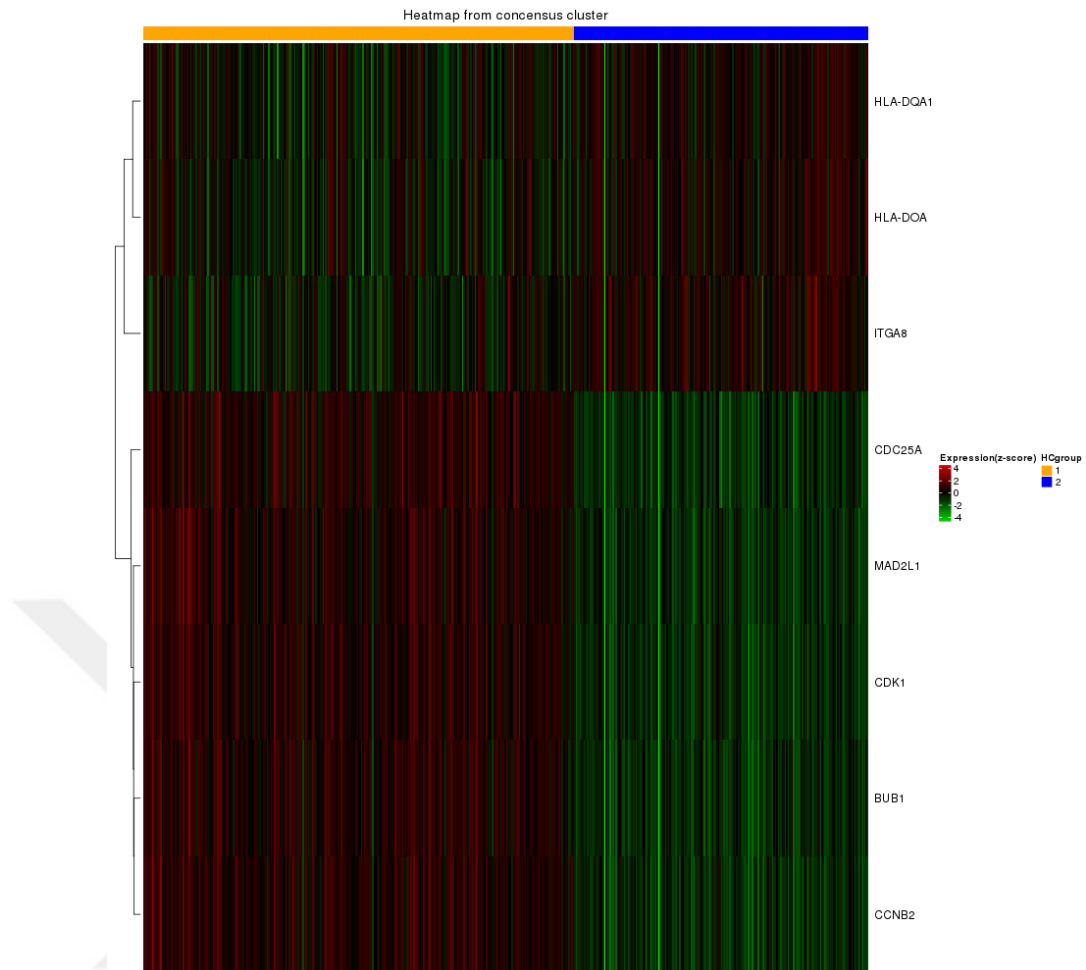


Figure 3.46. Heatmap of significant DEGs from multivariate CPHR for 510 LUAD patients

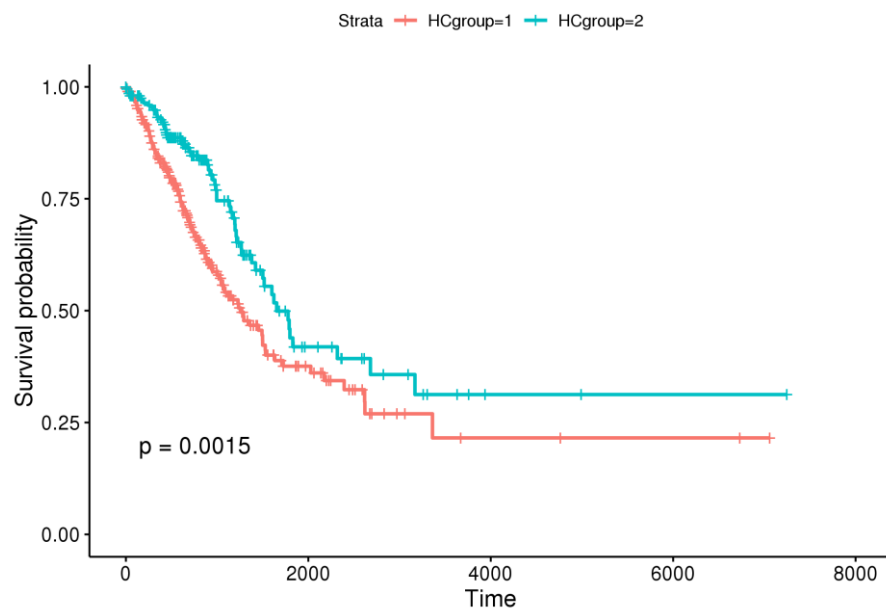


Figure 3.47. Overall survival plot of hierarchical clusters of 510 LUAD patients

4. DISCUSSION

In this study, we performed an integrative analysis by using level-3 data of SNVs, CNVs and transcriptome profiling of TCGA LUAD patients. We aimed to identify genomic and transcriptomic abnormalities that affect the overall survival and might be used for therapeutic predictions and personalized cancer therapy. We determined significantly mutated genes; amplified and deleted genes; and differentially expressed genes (DEGs) and their active subnetworks, by using R packages. Then we performed univariate and multivariate Cox Proportional Hazards Regression (CPHR) analysis to identify impactful genes on patients' survival time.

We identified that KRAS and EGFR oncogenes with TP53, STK11, RB1, and MGA tumor suppressors were mutated significantly in tumor samples of 55 LUAD patients. These genes were determined as candidate driver genes although they have a low number of mutations in samples of the cohort. KRAS and EGFR work in important signaling pathways playing roles in the regulation of cell proliferation and transcription of tumor suppressors. TP53, STK11, and RB1 are tumor suppressors playing important roles in cell metabolism, apoptosis, DNA damage response and regulation of cell division. Although KRAS, EGFR, TP53, STK11, and RB1 are well-known cancer-related genes, loss-of-function MGA mutations with MYC amplification have been newly described (The Cancer Genome Atlas Research Network, 2014). MGA encodes MAX gene-associated protein which is MYC-interacting transcription factor and antagonizes the transcriptional regulation of MYC involved in cancer processes (Romero, 2014).

We also identified 82 significant mutated genes, most of which are classified as oncogene (OG) or tumor suppressor gene (TSG), in tumor samples of 510 LUAD patients. EGFR, KRAS, TP53, STK11, RB1 and MGA genes are also in this list with

other 76 significant mutated genes including BRAF, PIK3CA, ATM, ERBB2, MAP2K1, SMAD4 and CDKN2A which are related with lung cancer and listed in cancer gene census. The other known cancer-related genes are NF1, SETD2, CTNNB1, ARID1A, ARID2, SMARCA4, NFE2L2, KDM5C, APC, GNAS, FBXW7, ATRX, CDKN1B, PBRM1, AKT1, RAF1, MAP2K4, B2M, BAP1.

We determined 5703 and 25177 gene aberrations (CNVs) in tumor samples of 55 and 510 LUAD patients respectively. This quite high number of aberrant genes play role in not only cancer related pathways such as Small Cell Lung Cancer Signaling, Role of p14/p19ARF in Tumor Suppression, NF- κ B Signaling and Molecular Mechanisms of Cancer, but also mostly in immune system pathways such as Complement System, Dendritic Cell Maturation, T Helper Cell Differentiation, IL-12 Signaling, Cytokine Signaling and Natural Killer Cell Signaling. The aberrations in immune system pathways suggest that tumor-immune system relationships are the key mechanisms in cancer cell survival and tumor growth.

We determined differentially expressed genes with adjusted p-values (q-values) in tumor samples of LUAD patients. In tumor samples of 55 and 510 LUAD patients, respectively 3591 and 3972 genes are dysregulated more than two-fold with 0.05 q-value significance. The most significant dysregulated pathways of DEGs are including Inhibition of Matrix Metalloproteases, EIF2 Signaling, Cell Cycle: G2/M DNA Damage Checkpoint Regulation, cAMP-mediated Signaling, Cell Cycle Control of Chromosomal Replication, Colorectal cancer metastasis signaling and Human embryonic stem cell pluripotency with upregulation of Cell Cycle, Mitosis, DNA Damage, Hereditary Breast cancer signaling, Bladder cancer signaling, GADD45 signaling and Metabolic pathways such as Glycolysis and HIF1 Signaling as expected to be highly upregulated in cancer cells. But there is also dysregulation in immune system pathways such as Granulocyte/Agranulocyte adhesion and diapedesis, Leukocyte extravasation signaling, Complement system, and IL-8 Signaling. The pathways consisting of downregulated genes are cAMP-mediated signaling, G-protein

coupled receptor signaling, G α i signaling, VEGF Family Ligand-Receptor Interactions, Human Embryonic Stem Cell Pluripotency, Protein Kinase A Signaling and other immune system pathways such as Complement system, Granulocyte/Agranulocyte adhesion and diapedesis, Dendritic cell maturation and T Helper cell differentiation. cAMP signaling is one of the important signaling pathways playing roles in metabolism, differentiation, proliferation, apoptosis. G-protein coupled receptor signaling, VEGF Family, Human Embryonic Stem Cell Pluripotency and Protein Kinase A Signaling are highly related to cancer. Another interesting downregulated pathway is G α i signaling which is a chemokine signaling pathway works through MAPK signaling pathway for cellular growth and differentiation, cell survival, migration, apoptosis, and metastasis. The dysregulation of these pathways may be directly or indirectly related to tumorigenesis, tumor growth or other mechanisms resulting in hallmarks of cancer.

Active subnetworks of differentially expressed genes in tumor samples of both 55 and 510 LUAD patients were determined. In tumor samples of 55 LUAD patients, 35 subnetworks including 173 genes were determined while 15 subnetworks including 66 genes were determined in tumor samples of 510 LUAD patients. The remarkable important genes are WNT genes, AKT3, FGFR4, PIK3R5 and PIK3R1 in active subnetworks. WNT pathway and FGF (Fibroblast Growth Factor) pathway are playing a role in cell proliferation, survival, cell differentiation, cell migration, angiogenesis and organogenesis during development and highly related to lung cancer. PIK3R genes are regulatory subunits of PI3K complex and AKT3 is one of the AKT family members, involved in tumorigenesis and modulations of several tumors. The pathways of subnetwork genes are mostly cancer-related pathways such as Small cell lung cancer, Melanoma, Glioma, Colorectal cancer, Chronic myeloid leukemia, Basal cell carcinoma, Prostate cancer, Pancreatic cancer, Bladder cancer, Cell cycle, Apoptosis, ERBB signaling, JAK-STAT signaling, and MAPK signaling pathways. The other pathways such as Regulation of actin cytoskeleton, Focal

adhesion, ECM receptor interaction, Cell adhesion molecules, Cell communication, Cytokine-cytokine receptor interaction, and B cell receptor signaling pathways are most probably related with metastasis and immune system regulation of tumors.

We integrated significant SNVs, CNVs, and DEGs from active subnetworks by performing multivariate Cox Proportional Hazards Regression (CPHR) analysis after univariate CPHR. We determined a set of genes including 4 mutated genes and 25 DEGs for 55 LUAD patients. When we drew heatmap for expression levels of these 25 DEGs, a dichotomous pattern was formed. Then we clustered the cohort into two and performed survival analysis resulted in the significant divergence of survival probability of two patient clusters. The mutated genes RB1, STK11, EGFR, and KRAS are well known and characterized cancer-related genes. The 25 DEGs mostly play role in cancer pathways such as Non-small Cell Lung Cancer Signaling, Glioma Signaling, Glioblastoma Multiforme Signaling, Colorectal Cancer Metastasis Signaling, Ovarian Cancer Signaling, Molecular Mechanisms of Cancer, Breast Cancer Regulation by Stathmin1, HER2 Signaling in Breast Cancer and highly lung cancer related pathways such as G Beta Gamma Signaling, Phospholipase C Signaling, IL-8 Signaling, Thrombin Signaling, G-Protein Coupled Receptor Signaling, UVC-induced MAPK Signaling and ERK/MAPK Signaling. One of active subnetwork DEGs, PRKCE (Protein kinase C epsilon type), is highly cancer-related serine/threonine-protein kinase which plays roles in the regulation of cell adhesion, cell migration, cancer cell invasion, cell cycle, apoptosis, and immune response and positive regulation of MAPK signaling (Sharma, 2007). WNT3A and GPC3 playing roles in canonical Wnt signaling and non-canonical Wnt signaling respectively, which may regulate tumor growth and migration in liver cancer (Gao, 2015; Li, 2019). GNG7 is a subunit of Guanine nucleotide-binding protein and ARHGEF6 is a guanine nucleotide exchange factor (GEF) and activates the Ras-like family of Rho proteins which acts at transmembrane signaling systems. PDE3B is Cyclic nucleotide phosphodiesterase which regulates the cAMP binding of RAPGEF3 which is a

Guanine nucleotide exchange factor (GEF) activated by binding cAMP and works for RAP1A and RAP2A small GTPases. The binding of PDE3B, APGEF3, and PIK3R6, activates the PI3K gamma complex and involved in angiogenesis (Wilson, 2011). The other angiogenesis regulators are VEGFD which is active in angiogenesis and endothelial cell growth, stimulating their proliferation and migration (Leppänen, 2011) and ANGPT1 (Angiopoietin-1) which is a key protein in the regulation of angiogenesis, endothelial cell survival, proliferation, migration, adhesion and cell spreading, with reorganization of the actin cytoskeleton in development during embryogenesis (Fukuhara, 2008). C2, CD244, CD74, IL11RA, and KLRD1 are key regulators in the complement system, NK cell cytotoxicity and antigen processing for the immune system. COL4A3, COL6A5, ROBO2, and PTGER4 receptors play a role in cell structure and cell migration. LEPR, LPL, and PLA2G3 are cell metabolism genes which may role in cancer metabolism, too. SCN4B is a metastasis-suppressor gene reducing cancer cell migration and tumor progression in breast cancer (Bon, 2016). KLF2 transcription factor is a putative tumor suppressor gene in non-small cell lung cancer (Jiang, 2017). The other genes CACNA1D and NR3C2 are not directly related to lung cancer but they were determined as impactful on overall survival of cancer patients.

Univariate Cox Proportional Hazards Regression (CPHR) analysis was performed for the clinical features such as gender, age, tumor stage and years of smoking of 510 LUAD patients. Because only the tumor stage can affect survival time significantly, tumor stage was put in multivariate CPHR analysis. Univariate CPHR analysis was also performed for significant SNVs, CNVs and active subnetwork DEGs of 510 LUAD patients, which were identified at previous steps. Then multivariate CPHR was performed for significant SNVs, CNVs and subnetwork DEGs from univariate CPHR. Multivariate CPHR results include 303 CNVs, 41 SNVs, 8 active subnetwork DEGs and tumor stage as impactful variates on overall survival of LUAD patients.

When we look at the impactful genes of CNVs, there are mostly pseudogenes, noncoding RNAs including long noncoding RNAs such as amplified ELDR (EGFR Long Non-Coding Downstream RNA), antisense RNAs such as deleted CDKN2B-AS1 which is effector of cell cycle arrest and miRNAs such as amplified miR-6837, amplified miR-1200 which showed negative correlation to the tumor stage in neuroendocrine tumors of the lung (Mairinger, 2014) and may target CD3 which is vital for T-cell development and cell recognition through TCR (T-cell receptor) - CD3 complex (Asghari Alashti, 2013) and amplified miR-5091 which was upregulated in LUAD samples (Hu, 2018). Other salient gene aberration are including amplifications of BCL3 (B-cell lymphoma 3-encoded protein), AKT2, FGFR1OP2 (Fibroblast growth factor receptor 1 oncogene partner 2), ABCA13 (ATP Binding Cassette Subfamily A Member 13), MUC22 (Mucin 22), IGF2BP3 (Insulin-like growth factor 2 mRNA-binding protein 3), IGFBP3 (Insulin-Like Growth Factor Binding Protein 3), HDAC9 (Histone Deacetylase 9), CREB5 (Cyclic AMP-responsive element-binding protein 5), ELMO1 (Engulfment and cell motility protein 1), MDM2 (E3 ubiquitin-protein ligase Mdm2), CDK13 (Cyclin-dependent kinase 13), MSRB3, PSPH and RAB3IP genes. BCL3 is a proto-oncogene candidate and its transcription is regulated by NF-KB signaling in colorectal cancer (Tao, 2018) and BCL3 gene mutation is strongly associated with a five-year survival of non-small-cell lung cancer patients (Dimitrakopoulos, 2015) but little is known about the possible role of BCL3 in lung carcinogenesis. AKT2 is one of AKT serine/threonine-protein kinases and AKT2 is playing a role in non-small cell lung cancer cell survival, growth, and migration (Lee, 2011). ABCA13 is a transmembrane transporter which is responsible for energy-dependent transport of substrate molecules and is a candidate marker for advanced staging in gastric cancer (Araújo, 2016) and poor survival in metastatic ovarian serous carcinoma (Nymoen, 2015). MUC22 was downregulated in tyrosine kinase inhibitor resistance lung cancer cell line (Wu, 2016). IGF2BP3 induces lung cancer development by reducing p53 stability and may serve as a potential diagnostic

marker for lung cancer (Zhao, 2017). On the other hand, IGFBP3 attenuates lung cancer development and cell growth through IGF1 Signaling (Wang, 2017). Overexpression of HDAC9 promotes cell division, tumor progression and poor prognosis in breast cancer, osteosarcoma and lung cancer (Zhao, 2015; Okudela, 2013; Huang, 2018). Histone Deacetylase Inhibitors are used to target HDAC proteins in lung cancer (Witta, 2012) and can reverse paclitaxel resistance in non-small cell lung cancer (Wang, 2016). CREB5 induces cell invasion and cause poor prognosis in ovarian cancer (He, 2017). ELMO1 also plays a role in cell invasion through phosphorylation by Axl receptor tyrosine kinase in non-small cell lung cancer (Abu-Thuraia, 2015; Zhang, 2018). MDM2 is responsible for ubiquitination of p53 for degradation by the proteasome, results in cell cycle arrest and apoptosis. Higher expression of MDM2 is correlated with poor clinical outcome of patients (Javid, 2015) and inhibition of MDM2-p53 interaction reduces tumor growth in Patient-Derived Non-Small Cell Lung Cancer Xenograft Models (Hai, 2015). CDK13 is responsible for RNA splicing and transcription elongation; and its over-expression has been demonstrated in lung adenocarcinoma (Bidkhori, 2013). MSRB3 and PSPH are playing roles in amino acid metabolism while RAB3IP is related with cell autophagy signaling but they have not been previously implicated in lung cancer.

Multivariate CPHR analysis could detect known lung cancer related mutated genes such as CDKN1B, MAP2K4, CDKN2A, STK31, BRAF and KRAS with expressions of 8 active subnetwork DEGs which are HLA-DOA, CDC25A, HLA-DQA1, BUB1, ITGA8, CCNB2, MAD2L1, CDK1. HLA-DOA and HLA-DQA1 play a role in antigen presentation process, and HLA-DQA1 is associated with Lung adenocarcinoma and Lung squamous cell carcinoma (Kohno, 2010; Kohno, 2011). Overexpression of cdc25A is frequent and BUB1 mitotic checkpoint serine/threonine kinase is a potential diagnostic and prognostic biomarker in NSCLC (Wu, 1998; Huang, 2018). BUB1 is also prognostic protein in The Human Protein Atlas (Uhlén, 2005) and high expression of BUB1 is unfavorable in lung cancer. Highly Expressed

ITGA8 increases epithelial-mesenchymal transition like features in Multiple Myeloma (Ryu, 2016) and is hypomethylated with decreased expression level in lung cancer (Mullapudi, 2015). CCNB2 is responsible for cell cycle control and overexpression of CCNB2 is associated with clinical progression and poor prognosis in NSCLC (Qian, 2015). CCNB2 is also prognostic protein in The Human Protein Atlas and high expression is unfavorable in lung cancer. MAD2L1 which is prognostic protein in The Human Protein Atlas and high expression is unfavorable in lung cancer, is playing role in spindle checkpoint and genetic variants in this gene cause susceptibility to lung cancer (Guo, 2010). Iron-dependent CDK1 activity induces lung carcinogenesis (Kuang, 2019) and downregulation of CDK1 causes poor survival and chemotherapeutic resistance in lung cancer (Zhang, 2011). CDK1 is prognostic in The Human Protein Atlas, high expression is unfavorable in lung cancer, too. CCNB2 (Cyclin B2), CDK1 and BUB1 work in cell cycle process (Bidkhori, 2013); CCNB1, CCNB2, CDK1, and MAD2L1 are significantly upregulated in SCLC samples (Ni, 2018); CDK1 and MAD2L1 are correlated with a high risk of cancer recurrence and poor survival in lung adenocarcinoma (Shi, 2016). Tumor stage is also impactful with other variates on overall survival of LUAD patients, that needs to be considered in patient-based lung cancer therapy.

5. CONCLUSION

In this study, we analyzed significant SNVs, CNVs, and DEGs in active subnetworks, which have an impact on overall survival of TCGA LUAD patients. We determined that the SNVs, CNVs, and expression of active subnetwork DEGs are impactful on the survival of patients, showed as a significant divergence of survival probability of patient clusters based on heatmap and hierarchical clustering. The impactful SNVs, CNVs and active subnetwork DEGs include both well-known lung cancer-related genes and recently identified lung cancer-related genes. However, there are some genes which have not been demonstrated for evidence of direct interaction between these genes and lung adenocarcinoma, yet. These impactful genes determined by Cox Proportional Hazard Regression Analysis, are candidates to be used as a molecular signature as diagnostic and prognostic biomarkers for this deadly disease.

REFERENCES

- Abu-Thuraia, A., Gauthier, R., Chidiac, R., Fukui, Y., Screaton, R. A., Gratton, J. P., & Côté, J. F. (2015). Axl phosphorylates Elmo scaffold proteins to promote Rac activation and cell invasion. *Molecular and Cellular Biology*, 35(1), 76–87.
- Aburatani, H., Enomoto, M., Watanabe, H., Soda, M., Choi, Y. L., Sugiyama, Y., Takada, S. (2007). Identification of the transforming EML4–ALK fusion gene in non-small-cell lung cancer. *Nature*, 448(7153), 561–566.
- Aramburu, A., Zudaire, I., Pajares, M. J., Agorreta, J., Orta, A., Lozano, M. D., Montuenga, L. M. (2015). Combined clinical and genomic signatures for the prognosis of early stage non-small cell lung cancer based on gene copy number alterations. *BMC Genomics*, 16(1), 1–10.
- Araújo, T. M., Seabra, A. D., Lima, E. M., Assumpção, P. P., Montenegro, R. C., Demachki, S., Khayat, A. S. (2016). Recurrent amplification of RTEL1 and ABCA13 and its synergistic effect associated with clinicopathological data of gastric adenocarcinoma. *Molecular Cytogenetics*, 9(1), 52.
- Asghari Alashti, F., & Minuchehr, Z. (2013). MiRNAs which target CD3 subunits could be potential biomarkers for cancers. *PloS One*, 8(11), e78790.
- Berger, A. H., Brooks, A. N., Wu, X., Shrestha, Y., Chouinard, C., Piccioni, F., Boehm, J. S. (2016). High-throughput Phenotyping of Lung Cancer Somatic Mutations. *Cancer Cell*, 30(2), 214–228.
- Bidkhor G, Narimani Z, Hosseini Ashtiani S, Moeini A, Nowzari-Dalini A, et al. (2013) Reconstruction of an Integrated Genome-Scale Co-Expression Network Reveals Key Modules Involved in Lung Adenocarcinoma. *PLOS ONE* 8(7): e67552.
- Bromberg, Y. (2013). Chapter 15: Disease Gene Prioritization. *PLoS Computational Biology*, 9(4).
- Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-Wide Association Studies. *PLoS Computational Biology*, 8(12).
- Cardoso, F., van't Veer, L. J., Bogaerts, J., Slaets, L., Viale, G., Delaloge, S., Piccart, M. (2016). 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. *New England Journal of Medicine*, 375(8), 717–729.
- Chang, K., Creighton, C. J., Davis, C., Donehower, L., Drummond, J., Wheeler, D., Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10), 1113–1120.

- Deng, Z. min, Liu, L., Qiu, W. hai, Zhang, Y. qun, Zhong, H. yan, Liao, P., & Wu, Y. hong. (2017). Analysis of genomic variation in lung adenocarcinoma patients revealed the critical role of PI3K complex. *PeerJ*, 5, e3216.
- Dimitrakopoulos, F.-I. D., Antonacopoulou, A. G., Kottorou, A., Marousi, S., Koukourikou, I., Kalofonou, M., Kalofonos, H. P. (2015). Variant of BCL3 gene is strongly associated with five-year survival of non-small-cell lung cancer patients. *Lung Cancer*, 89(3), 311–319.
- Ding, L., Getz, G., Wheeler, D. A., Mardis, E. R., McLellan, M. D., Cibulskis, K., Wilson, R. K. (2008). Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, 455(7216), 1069–1075.
- Govindan, R., Ding, L., Griffith, M., Subramanian, J., Dees, N. D., Kanchi, K. L., Wilson, R. K. (2012). Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell*, 150(6), 1121–1134.
- Guo, Y., Zhang, X., Yang, M., Miao, X., Shi, Y., Yao, J., Lin, D. (2010). Functional evaluation of missense variations in the human MAD1L1 and MAD2L1 genes and their impact on susceptibility to lung cancer. *Journal of Medical Genetics*, 47(9), 616 LP – 622.
- Hai, J., Sakashita, S., Allo, G., Ludkovski, O., Ng, C., Shepherd, F. A., & Tsao, M.-S. (2015). Inhibiting MDM2-p53 Interaction Suppresses Tumor Growth in Patient-Derived Non-Small Cell Lung Cancer Xenograft Models. *Journal of Thoracic Oncology*, 10(8), 1172–1180.
- Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell*, 144(5), 646–674.
- Hanash, S. (2004). Integrated global profiling of cancer. *Nature Reviews Cancer*, 4(8), 638–644.
- He, S., Deng, Y., Liao, Y., Li, X., Liu, J., & Yao, S. (2017). CREB5 promotes tumor cell invasion and correlates with poor prognosis in epithelial ovarian cancer. *Oncology Letters*, 14(6), 8156–8161.
- Hicks, C., Asfour, R., Pannuti, A., & Miele, L. (2011). An integrative genomics approach to biomarker discovery in breast cancer. *Cancer Informatics*, 10, 185–204.
- Hicks, C., Koganti, T., Giri, S., Tekere, M., Ramani, R., Sitthi-amorn, J., & Vijayakumar, S. (2014). Integrative Genomic Analysis for the Discovery of Biomarkers in Prostate Cancer. *Biomarker Insights*, 9, 39–51.
- Hu, Y., Dingerdissen, H., Gupta, S., Kahsay, R., Shanker, V., Wan, Q., Mazumder, R. (2018). Identification of key differentially expressed MicroRNAs in cancer patients through pan-cancer analysis. *Computers in Biology and Medicine*, 103, 183–197.

- Huang, R., & Huang, R. (2018). Identification of potential diagnostic and prognostic biomarkers in non-small cell lung cancer based on microarray data. *Oncology Letters*, 15, 6436-6442.
- Huang, T., Yang, J., & Cai, Y. D. (2015). Novel candidate key drivers in the integrative network of genes, microRNAs, methylations, and copy number variations in squamous cell lung carcinoma. *BioMed Research International*, 2015.
- Huang, Y., Jian, W., Zhao, J., & Wang, G. (2018). Overexpression of HDAC9 is associated with poor prognosis and tumor progression of breast cancer in Chinese females. *OncoTargets and Therapy*, 11, 2177–2184.
- Imielinski, M., Berger, A. H., Hammerman, P. S., Hernandez, B., Pugh, T. J., Hodis, E., Meyerson, M. (2012). Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell*, 150(6), 1107–1120.
- Javid, J., Mir, R., Julka, P. K., Ray, P. C., & Saxena, A. (2015). Association of p53 and mdm2 in the development and progression of non-small cell lung cancer. *Tumor Biology*, 36(7), 5425–5432.
- Ju, Y. S., Lee, W.-C., Shin, J.-Y., Lee, S., Bleazard, T., Won, J.-K., Seo, J.-S. (2012). A transforming KIF5B and RET gene fusion in lung adenocarcinoma revealed from whole-genome and transcriptome sequencing. *Genome Research*, 22(3), 436–445.
- Kan, Z., Jaiswal, B. S., Stinson, J., Janakiraman, V., Bhatt, D., Stern, H. M., Seshagiri, S. (2010). Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature*, 466(7308), 869–873.
- Kaplan, E. L., & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282), 457–481.
- Kassambara A., Kosinski M., Biecek P., Fabian S. (2019). Drawing Survival Curves using 'ggplot2', version 0.4.4, <https://CRAN.R-project.org/package=survminer>
- Kohno, T., Kunitoh, H., Shimada, Y., Shiraishi, K., Ishii, Y., Goto, K., Yokota, J. (2010). Individuals susceptible to lung adenocarcinoma defined by combined HLA-DQA1 and TERT genotypes. *Carcinogenesis*, 31(5), 834–841.
- Kohno, T., Kunitoh, H., Mimaki, S., Shiraishi, K., Kuchiba, A., Yamamoto, S., & Yokota, J. (2011). Contribution of the TP53, OGG1, CHRNA3, and HLA-DQA1 Genes to the Risk for Lung Squamous Cell Carcinoma. *Journal of Thoracic Oncology*, 6(4), 813–817.
- Kohno, T., Ichikawa, H., Totoki, Y., Yasuda, K., Hiramoto, M., Nammo, T., Shibata, T. (2012). KIF5B-RET fusions in lung adenocarcinoma. *Nature Medicine*, 18(3), 375–377.
- Kristensen, V. N., Lingjærde, O. C., Russnes, H. G., Vollen, H. K. M., Frigessi, A., & Børresen-Dale, A. L. (2014). Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer*, 14(5), 299–313.

- Kuang, Y., Guo, W., Ling, J., Xu, D., Liao, Y., Zhao, H., Wang, Q. (2019). Iron-dependent CDK1 activity promotes lung carcinogenesis via activation of the GP130/STAT3 signaling pathway. *Cell Death & Disease*, 10(4), 297.
- Laubenbacher, R., Hower, V., Jarrah, A., Torti, S. V., Shulaev, V., Mendes, P., Akman, S. (2009). A systems biology view of cancer. *Biochimica et Biophysica Acta - Reviews on Cancer*, 1796(2), 129–139.
- Lee, M. W., Kim, D. S., Lee, J. H., Lee, B. S., Lee, S. H., Jung, H. L., Sung, K. W., Kim, H. T., Yoo, K. H. and Koo, H. H. (2011), Roles of AKT1 and AKT2 in non-small cell lung cancer cell survival, growth, and migration. *Cancer Science*, 102: 1822-1828.
- Lindeman, N. I., Cagle, P. T., Beasley, M. B., Chitale, D. A., Dacic, S., Giaccone, G., Ladanyi, M. (2013). Molecular Testing Guideline for Selection of Lung Cancer Patients for EGFR and ALK Tyrosine Kinase Inhibitors: Guideline from the College of American Pathologists, International Association for the Study of Lung Cancer, and Association for Molecular Patho. *Journal of Thoracic Oncology*, 8(7), 823–859.
- Liu, P., Morrison, C., Wang, L., Xiong, D., Vedell, P., Cui, P., You, M. (2012). Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing. *Carcinogenesis*, 33(7), 1270–1276.
- Mairinger, F. D., Ting, S., Werner, R., Walter, R. F. H., Hager, T., Vollbrecht, C., Wohlschlaeger, J. (2014). Different micro-RNA expression profiles distinguish subtypes of neuroendocrine tumors of the lung: results of a profiling study. *Modern Pathology*, 27, 1632.
- Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, 50(3), 163–170.
- Mullapudi, N., Ye, B., Suzuki, M., Fazzari, M., Han, W., Shi, M. K., Spivack, S. D. (2015). Genome Wide Methylome Alterations in Lung Cancer. *PloS One*, 10(12), e0143826.
- Ni, Z., Wang, X., Zhang, T., Li, L., & Li, J. (2018). Comprehensive analysis of differential expression profiles reveals potential biomarkers associated with the cell cycle and regulated by p53 in human small cell lung cancer. *Experimental and Therapeutic Medicine*, 15, 3273-3282.
- Nymoen, D. A., Holth, A., Hetland Falkenthal, T. E., Tropé, C. G., & Davidson, B. (2015). CIAPIN1 and ABCA13 are markers of poor survival in metastatic ovarian serous carcinoma. *Molecular Cancer*, 14(1), 44.
- Okudela, K., Mitsui, H., Suzuki, T., Woo, T., Tateishi, Y., Umeda, S., Ohashi, K. (2013). Expression of HDAC9 in lung cancer--potential role in lung carcinogenesis. *International Journal of Clinical and Experimental Pathology*, 7(1), 213–220.

- Qian, X., Song, X., He, Y., Yang, Z., Sun, T., Wang, J., You, C. (2015). CCNB2 overexpression is a poor prognostic biomarker in Chinese NSCLC patients. *Biomedicine & Pharmacotherapy*, 74, 222–227.
- Ryu, J., Koh, Y., Park, H., Kim, D. Y., Kim, D. C., Byun, J. M., Yoon, S. S. (2016). Highly Expressed Integrin- $\alpha 8$ Induces Epithelial to Mesenchymal Transition-Like Features in Multiple Myeloma with Early Relapse. *Molecules and Cells*, 39(12), 898–908.
- Shi, J., Hua, X., Zhu, B., Ravichandran, S., Wang, M., Nguyen, C., Landi, M. T. (2016). Somatic Genomics and Clinical Features of Lung Adenocarcinoma: A Retrospective Study. *PLoS Medicine*, 13(12), 1–24.
- Shi, Y. X., Zhu, T., Zou, T., Zhuo, W., Chen, Y. X., Huang, M. S., Liu, Z. Q. (2016). Prognostic and predictive values of CDK1 and MAD2L1 in lung adenocarcinoma. *Oncotarget*, 7(51), 85235–85243.
- Tao, Y., Liu, Z., Hou, Y., Wang, S., Liu, S., Jiang, Y., Zhang, X. (2018). Alternative NF- κ B signaling promotes colorectal tumorigenesis through transcriptionally upregulating Bcl-3. *Oncogene*, 37(44), 5887–5900.
- The Cancer Genome Atlas Research Network. (2014). Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511), 543–550.
- The International Agency for Research on Cancer (IARC). Latest global cancer data: Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018. 12 September 2018. PRESS RELEASE No 263. *World Health Organization*. Geneva, Switzerland.
- Therneau T. M., Grambsch P. M. (2000). Modeling Survival Data: Extending the Cox Model. *Springer*, New York. ISBN 0-387-98784-3.
- Therneau T. (2019). A Package for Survival Analysis, version 2.44-1.1, <https://CRAN.R-project.org/package=survival>.
- Thorgeirsson SS, Grisham JW (2002). Molecular pathogenesis of human hepatocellular carcinoma. *Nature Genetics*, 31:339-346.
- Togashi, Y., Lim Choi, Y., Nakagawa, K., Hamanaka, W., Soda, M., Satoh, Y., Uehara, H. (2012). RET, ROS1 and ALK fusions in lung cancer. *Nature Medicine*, 18(3), 378–381.
- Travis, W. D. (2011). Pathology of Lung Cancer. *Clinics in Chest Medicine*, 32(4), 669–692.
- Uhlén, M., Björling, E., Agaton, C., Szigartyo, C. A.-K., Amini, B., Andersen, E., Pontén, F. (2005). A Human Protein Atlas for Normal and Cancer Tissues Based on Antibody Proteomics. *Molecular & Cellular Proteomics*, 4(12), 1920 LP – 1932.

- van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. a. M., Mao, M., Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Letters to Nature*, 415(345), 530–536.
- Vazquez, M., de la Torre, V., & Valencia, A. (2012). Chapter 14: Cancer Genome Analysis. *PLoS Computational Biology*, 8(12), 1–9.
- Wang, L., Li, H., Ren, Y., Zou, S., Fang, W., Jiang, X., Wu, C. (2016). Targeting HDAC with a novel inhibitor effectively reverses paclitaxel resistance in non-small cell lung cancer via multiple mechanisms. *Cell Death & Disease*, 7(1), e2063.
- Wang, Y. A., Sun, Y., Palmer, J. D., Solomides, C., Huang, L.-C., Shyr, Y., Lu, B. (2017). IGFBP3 Modulates Lung Tumorigenesis and Cell Growth through IGF1 Signaling. *Molecular Cancer Research*, molcanres.0390.2016.
- Weir, B. A., Woo, M. S., Getz, G., Perner, S., Ding, L., Beroukhi, R., Meyerson, M. (2007). Characterizing the cancer genome in lung adenocarcinoma. *Nature*, 450(7171), 893–898.
- Witta, S. (2012). Histone Deacetylase Inhibitors in Non-Small-Cell Lung Cancer. *Journal of Thoracic Oncology*, 7(12), S404–S406.
- Wu, W., Fan, Y.-H., Kemp, B. L., Walsh, G., & Mao, L. (1998). Overexpression of cdc25A and cdc25B Is Frequent in Primary Non-Small Cell Lung Cancer but Is Not Associated with Overexpression of c-myc. *Cancer Research*, 58(18), 4082 LP – 4085.
- Wu, Y., Yu, D., Hu, Y., Yan, D., Chen, X., Cao, H. Feng, J. (2016). Genome-wide profiling of long non-coding RNA expression patterns in the EGFR-TKI resistance of lung adenocarcinoma by microarray. *Oncology Reports*, 35, 3371-3386.
- Zack, T. I., Schumacher, S. E., Carter, S. L., Cherniack, A. D., Saksena, G., Tabak, B., Beroukhi, R. (2013). Pan-cancer patterns of somatic copy number alteration. *Nature Genetics*, 45(10), 1134–1140.
- Zhang, C., Elkahoul, A. G., Robertson, M., Gills, J. J., Tsurutani, J., Shih, J. H., Dennis, P. A. (2011). Loss of cytoplasmic CDK1 predicts poor survival in human lung cancer and confers chemotherapeutic resistance. *PloS One*, 6(8), e23849.
- Zhang, D. Y., Ye, F., Gao, L., Liu, X., Zhao, X., Che, Y., Lee, P. (2009). Proteomics, pathway array and signaling network-based medicine in cancer. *Cell Division*, 4, 1–16.
- Zhang, J., Abrams, Z., Parvin, J. D., & Huang, K. (2016). Integrative analysis of somatic mutations and transcriptomic data to functionally stratify breast cancer patients. *BMC Genomics*, 17(Suppl 7).
- Zhang, G., Wang, M., Zhao, H., & Cui, W. (2018). Function of Axl receptor tyrosine kinase in non-small cell lung cancer. *Oncology Letters*, 15(3), 2726–2734.
- Zhang, Q., Burdette, J. E., & Wang, J. P. (2014). Integrative network analysis of TCGA data for ovarian cancer. *BMC Systems Biology*, 8, 1338.

Zhao, W., Lu, D., Liu, L., Cai, J., Zhou, Y., Yang, Y., Zhang, J. (2017). Insulin-like growth factor 2 mRNA binding protein 3 (IGF2BP3) promotes lung tumorigenesis via attenuating p53 stability. *Oncotarget*, 8(55), 93672–93687.

Zhao, Y. X., Wang, Y. S., Cai, Q. Q., Wang, J. Q., & Yao, W. T. (2015). Up-regulation of HDAC9 promotes cell proliferation through suppressing p53 transcription in osteosarcoma. *International Journal of Clinical and Experimental Medicine*, 8(7), 11818–11823.



APPENDIX A.

The R codes below can be found at <https://bitbucket.org/TalipZengin/tcga-pipeline> address.

Installation of R Packages

```
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")

BiocManager::install("DT", version = "3.8")
BiocManager::install("maftools", version = "3.8")
BiocManager::install("limma", version = "3.8")
BiocManager::install("edgeR", version = "3.8")
BiocManager::install("rafalib", version = "3.8")
BiocManager::install("gaia", version = "3.8")
BiocManager::install("TCGAbiolinks", version = "3.8")
BiocManager::install("statmod", version = "3.8")
BiocManager::install("data.table", version = "3.8")
BiocManager::install("tidyverse", version = "3.8")
BiocManager::install("DEsubs", version = "3.8")
BiocManager::install("biomaRt", version = "3.8")
BiocManager::install("rafalib", version = "3.8")
BiocManager::install("dnet", version = "3.8")
BiocManager::install("pathview", version = "3.8")

update.packages(ask = FALSE)

## SomInaClust Installation -----

if(!"plotrix"%in%installed.packages()) install.packages("plotrix")
if(!"foreach"%in%installed.packages()) install.packages("foreach")
if(!"doParallel"%in%installed.packages()) install.packages("doParallel")
if(!"parallel"%in%installed.packages()) install.packages("parallel")

install.packages("http://bioinformatics.intec.ugent.be/sominaclust/R/SomInaClust_1.0.0.tar.gz", type =
"source", repos = NULL)
```

APPENDIX A. (cont.)

R Codes for Analysis LUAD Patients

```
# TCGA ANALYSIS INTRO -----

## Determining the patients who have all types of data

library(TCGAbiolinks)
library(SummarizedExperiment)
library(dplyr)
library(DT)
library(maftools)

cancer <- "LUAD"

#GDCquery(project = paste0("TCGA-", cancer))

## SNP Samples and Patients -----

maf <- GDCquery_Maf(tumor = cancer, pipelines = "mutect2", save.csv = TRUE) %>% read.maf

cases_maf <- maf@clinical.data[["Tumor_Sample_Barcode"]]
write.table(cases_maf, file = paste0(cancer, "_cases_maf.tsv"), row.names=FALSE, na="NA", col.names=FALSE, sep="\t")

patients_maf <- unique(substring(cases_maf, 1, 12))
write.table(patients_maf, file = paste0(cancer, "_patients_maf.tsv"), row.names=FALSE, na="NA", col.names=FALSE, sep="\t")

## RNAseq Samples and Patients -----

query_exp <- GDCquery(project = paste0("TCGA-", cancer),
  data.category = "Transcriptome Profiling",
  data.type = "Gene Expression Quantification",
  workflow.type = "HTSeq - Counts")

cases_exp <- getResults(query_exp, cols="cases")
cases_exp2 <- TCGAquery_SampleTypes(cases_exp, c("NT", "TP"))
```

APPENDIX A. (cont.)

```
write.table(cases_exp2, file = paste0(cancer, "_cases_exp.tsv"),row.names=FALSE, na="NA",col.names=FALSE, sep="\t")

patients_exp <- unique(substring(cases_exp2, 1, 12))
write.table(patients_exp, file = paste0(cancer, "_patients_exp.tsv"), row.names=FALSE, na="NA",col.names=FALSE, sep="\t")

## CNV Samples and Patients -----

query_cnv <- GDCquery(project = paste0("TCGA-", cancer),
  data.category = "Copy Number Variation",
  data.type = "Masked Copy Number Segment")

cases_cnv <- getResults(query_cnv, cols="cases")
cases_cnv2 <- TCGAquery_SampleTypes(cases_cnv, c("NT","TP"))
write.table(cases_cnv2, file = paste0(cancer, "_cases_cnv.tsv"),row.names=FALSE, na="NA",col.names=FALSE, sep="\t")

patients_cnv <- unique(substring(cases_cnv2, 1, 12))
write.table(patients_cnv, file = paste0(cancer, "_patients_cnv.tsv"), row.names=FALSE, na="NA",col.names=FALSE, sep="\t")

## Clinical Data and Patients -----

clinic <- GDCquery_clinic(project = paste0("TCGA-", cancer),
  type = "clinical")

patients_clinic <- clinic$submitter_id
write.table(patients_clinic, file = paste0(cancer, "_patients_clinic.tsv"), row.names=FALSE, na="NA",col.names=FALSE, sep="\t")

## Intersection of Patients for All Type of Data -----

common_patients <- intersect(intersect(intersect(patients_maf,patients_exp),patients_cnv),patients_clinic)
write.table(common_patients, file = paste0(cancer, "_common_patients.tsv"), row.names=FALSE, na="NA",col.names=FALSE, sep="\t")
```

APPENDIX A. (cont.)

```
### SNP Analysis of TCGA-LUAD -----

library(TCGAbiolinks)
library(SummarizedExperiment)
library(dplyr)
library(DT)
library(maftools)

cancer <- "LUAD"

# These codes below are used to capture barcodes of the common patients who have both normal and tumor data
# from the file which was saved previously.
common_patients <- read.csv(file = paste0(cancer, "_common_patients.tsv"), header = FALSE, sep = "\t")
common_patients <- as.character(unlist(common_patients))

# GDCquery_Maf command from TCGAbiolinks package, downloads mutation data of all patients in the LUAD project
# in Mutation Annotation File (maf) format generated by the indicated pipeline type in the code.
maf <- GDCquery_Maf(tumor = cancer, pipelines = "mutect2", save.csv = TRUE) %>% read.maf

# These codes below determine the sample barcodes of common patients in mutation data.
cases_maf <- maf@clinical.data[["Tumor_Sample_Barcode"]]

cases_maf <- cases_maf[unique(substring(cases_maf, 1, 12)) %in% common_patients]
cases_maf <- as.character(unlist(cases_maf))
write.table(cases_maf, file = paste0(cancer, "_cases_maf.tsv"),
            row.names=FALSE, na="NA", col.names=FALSE, sep="\t")

# Filter TCGA Replicate Samples (cases_maf -> uniq_tsb_maf)
source("tcga_replicateFilter_snp.R")

# subsetMaf command from maftools package, extracts the mutation data of indicated barcodes,
# here the mutation data of common patients.
maf = subsetMaf(maf, tsb = uniq_tsb_maf, mafObj = TRUE)
save(maf, file=paste0(cancer, "_maf.RData"))
```

APPENDIX A. (cont.)

plotmafSummary command from TCGAbiolinks package summarizes the mutation data in graphics as a plot.

```
tiff(file = paste0(cancer, "_plotmafSummary.tif"), width = 7.5, height = 5.3, units = 'in', res = 300, compression = 'lzw')
```

```
plotmafSummary(maf = maf, rmOutlier = TRUE, addStat = 'median', dashboard = TRUE)
```

```
dev.off()
```

The command below summarizes the number, mean and median of mutation types in mutation data as a table.

```
write.table(maf@summary, file = paste0(cancer, "_maf_summary.tsv"), row.names=FALSE, na="NA", col.names=TRUE, sep="\t")
```

The command below summarizes the number of mutation types in each sample as a table.

```
write.table(maf@variant.type.summary, file = paste0(cancer, "_maf_variant.type.summary.tsv"), row.names=FALSE, na="NA", col.names=TRUE, sep="\t")
```

oncoplot command from TCGAbiolinks package summarizes the mutation types in gene-level as a plot.

```
tiff(file = paste0(cancer, "_oncoplot.tif"), width = 7.5, height = 5.3, units = 'in', res = 300, compression = 'lzw')
```

```
oncoplot(maf = maf, top = 10, removeNonMutated = TRUE)
```

```
dev.off()
```

titv and plotTiTv commands from TCGAbiolinks package summarize the substitution types in mutation data as a plot.

```
titv = titv(maf = maf, plot = FALSE, useSyn = TRUE)
```

```
tiff(file = paste0(cancer, "_plotTiTv.tif"), width = 7.5, height = 5.3, units = 'in', res = 300, compression = 'lzw')
```

```
plotTiTv(res = titv)
```

```
dev.off()
```

```
## COSMIC Reference Files -----
```

First download mutation reference file (CosmicMutantExport.tsv.gz) under "COSMIC Mutation Data" section and list of all cancer census genes (cancer_gene_census.csv) under "Cancer Gene Census" section file from the COSMIC database as a registered user.

APPENDIX A. (cont.)

```
# Open these files (cosmic_file specifies the path and the name of the downloaded file)
#cosmic_file <- R.utils::gunzip("CosmicMutantExport.tsv.gz")
#cosmic <- read.table(cosmic_file, header=TRUE, sep='\t', row.names=NULL, stringsAsFactors=FALSE, quote="")
#cosmic <- cosmic %>% select(Gene.name, Gene.CDS.length, Sample.name, Mutation.Description, Mutation.AA, Mutation.CDS, Mutation.genome.position) %>% rename(Mutation.GRCh37.genome.position = Mutation.genome.position)

## Driver Mutation Analysis by SomInaClust -----

library(SomInaClust)

# This command prepares the reference files which will be used during the determination of significant driver mutations.
#SomInaClust_ref(maf=cosmic,
#               database='COSMIC',
#               convert_cosmic_to_tcga=TRUE,
#               convert_genenames_to_HGNC=TRUE,
#               calculate_CDS=TRUE,
#               convert_cosmic_tcga_input=converttable_mutation_cosmic_tcga,
#               convert_genenames_HGNC_input=converttable_genenames_HGNC,
#               n_cores = parallel::detectCores(),
#               filename="SomInaClust_ref.txt")

# This command to load ref data
#ref_CDSlength <- read.table("SomInaClust_ref_CDSlength.txt",sep="\t",header=TRUE,
#                           row.names=1,stringsAsFactors = FALSE)
#ref_clusters <- read.table("SomInaClust_ref_clusters.txt",sep="\t",header=TRUE,
#                           stringsAsFactors = FALSE)
#ref_corr_factors <- read.table("SomInaClust_ref_corr_factors.txt",sep="\t",header=TRUE,
#                              row.names=1,stringsAsFactors = FALSE)

# This command
#cancer_gene_census <- read.csv("cancer_gene_census.csv", row.names=1)
#cancer_gene_census <- cancer_gene_census %>% select(Entrez.GeneId, Molecular.Genetics)

# Save
#save(ref_CDSlength, ref_clusters, ref_corr_factors, cancer_gene_census, file="SomInaClust_ref_v88_hg38.RData")
```

APPENDIX A. (cont.)

```
# Load cosmic GRCh38 data (v88, to be updated if necessary)
load("SomInaClust_ref_v88_hg38.RData")

# Add colname with CDS position
CDS_pos <- maf@data[, "CDS_position"] %>% unlist
CDS_pos <- gsub("-", ".*", "", CDS_pos)
maf@data[, "CDS_Position"] <- CDS_pos

# Add colname with c_position_WU to use calculate_CDS = TRUE option of SomInaClust_det command
c_position_WU <- paste0('c.', CDS_pos)
maf@data[, "c_position_WU"] <- c_position_WU

# This command
SomInaClust_result <- SomInaClust_det(maf = maf@data,
                                     database = "TCGA",
                                     save_output = TRUE,
                                     return_results = TRUE,
                                     convert_cosmic_to_tcga = FALSE,
                                     convert_genenames_to_HGNC = FALSE,
                                     calculate_CDS = TRUE,
                                     save_new_maf = TRUE,
                                     CDS_length = ref_CDSlength,
                                     cluster_matrix_input = ref_clusters,
                                     corr_factor_input = ref_corr_factors,
                                     filename = paste0(cancer, "_SomInaClust_results.txt"),
                                     n_cores = parallel::detectCores(),
                                     create_pyramidplot = TRUE,
                                     pyramidplot_n_genes = 40,
                                     create_summary_table = TRUE,
                                     CGC = cancer_gene_census,
                                     colname_CGC_class = "Molecular.Genetics")
```

APPENDIX A. (cont.)

```
## CNV Analysis of TCGA-LUAD

library(TCGAbiolinks)
library(gaia)
library(downloader)
library(readr)
library(data.table)
library(statmod)

load_cnv <- function (segmentation_matrix, markers_list, num_of_samples) {
  message("Loading Copy Number Data")
  chromosomes <- as.numeric(sort(unique(names(markers_list))))
  chromosomes <- chromosomes[which(!is.na(chromosomes))]
  aberration_kinds <- 1:length(unique(segmentation_matrix[,6]))
  names(aberration_kinds) <- sort(unique(segmentation_matrix[,6]))
  samples <- 1:num_of_samples
  if (!is.numeric(segmentation_matrix[, 1])) {
    sample_names <- unique(segmentation_matrix[, 1])
    for (i in 1:length(sample_names)) {
      segmentation_matrix[which(segmentation_matrix[, 1] ==
                                sample_names[i]), 1] <- i
    }
  }
  region_list <- list()
  for (k in 1:length(aberration_kinds)) {
    region_list[[aberration_kinds[k]]] <- list()
    for (i in 1:length(chromosomes)) {
      region_list[[aberration_kinds[k]]][[chromosomes[i]]] <- matrix(0,
                                                                    length(samples), ncol(markers_list[[chromosomes[i]]]))
      rownames(region_list[[aberration_kinds[k]]][[chromosomes[i]]]) <- samples
      colnames(region_list[[aberration_kinds[k]]][[chromosomes[i]]]) <- c(1:ncol(markers_list[[chromo
somes[i]]]))
    }
  }
  for (k in 1:length(aberration_kinds)) {
    ab_ids <- which(segmentation_matrix[, 6] == names(aberration_kinds[k]))
    tmp_matrix1 <- segmentation_matrix[ab_ids, ]
    if (class(tmp_matrix1) == "numeric") {
```

APPENDIX A. (cont.)

```
tmp_matrix1 <- t(as.matrix(tmp_matrix1))
}
for (i in 1:length(chromosomes)) {
  tmp_matrix2 <- tmp_matrix1[which(tmp_matrix1[, 2] ==
                                chromosomes[i]), ]
  if (class(tmp_matrix2) == "numeric") {
    tmp_matrix2 <- t(as.matrix(tmp_matrix2))
  }
  message(".", appendLF = FALSE)
  for (j in 1:length(samples)) {
    tmp_matrix3 <- tmp_matrix2[which(tmp_matrix2[,
                                    1] == samples[j]), ]
    if (class(tmp_matrix3) == "numeric") {
      tmp_matrix3 <- t(as.matrix(tmp_matrix3))
    }
    if (nrow(tmp_matrix3) > 0) {
      for (t in 1:nrow(tmp_matrix3)) {
        start_prob <- tmp_matrix3[t, 3]
        end_prob <- tmp_matrix3[t, 4]
        start_index <- which(markers_list[[chromosomes[i]]][1,] == start_prob)
        end_index <- which(markers_list[[chromosomes[i]]][2,] == end_prob)
        if ((length(start_index) > 0 ) & (length(end_index) > 0 )){
          region_list[[aberration_kinds[k]]][[chromosomes[i]]][samples[j],
                                              start_index:end_index] <- 1
        }
      }
    }
  }
}
}
}

message("\nDone")
names(region_list) <- names(aberration_kinds)
return(region_list)
}

cancer <- "LUAD"
common_patients <- read.csv(file = paste0(cancer, "_common_patients.tsv"), header = FALSE, sep =
"\t")
```

APPENDIX A. (cont.)

```
common_patients <- as.character(unlist(common_patients))

query_cnv <- GDCquery(project = paste0("TCGA-", cancer),
  data.category = "Copy Number Variation",
  data.type = "Masked Copy Number Segment",
  sample.type = "Primary solid Tumor",
  barcode = common_patients)

cases_cnv <- getResults(query_cnv, cols="cases")

# Filter TCGA Replicate Samples (cases_cnv -> uniq_tsb_cnv)
source("tcga_replicateFilter_cnv.R")

query_cnv <- GDCquery(project = paste0("TCGA-", cancer),
  data.category = "Copy Number Variation",
  data.type = "Masked Copy Number Segment",
  sample.type = "Primary solid Tumor",
  barcode = uniq_tsb_cnv)

GDCdownload(query_cnv, files.per.chunk = 100)

cnv_data <- GDCprepare(query_cnv)
save(cnv_data, file = paste0(cancer, "_cnv_data.rda"))

#####
download.file("https://gdc.cancer.gov/files/public/file/snp6.na35.liftoverhg38.txt.zip", destfile="snp6.n
a35.liftoverhg38.txt.zip")
file <- unzip("snp6.na35.liftoverhg38.txt.zip")

markersMatrix <- readr::read_tsv(file, col_names = TRUE, col_types = "ccn____", progress = TRUE)
colnames(markersMatrix) <- c("Probe.Name", "Chromosome", "Start")

save(markersMatrix, file = "markersMatrix.rda", compress = "xz")
#####

# Prepare CNV matrix
cnvMatrix <- cnv_data
```

APPENDIX A. (cont.)

```
# Add label (0 for loss, 1 for gain)
cnvMatrix <- cbind(cnvMatrix, Label=NA)
cnvMatrix[cnvMatrix[, "Segment_Mean"] < -0.3, "Label"] <- 0
cnvMatrix[cnvMatrix[, "Segment_Mean"] > 0.3, "Label"] <- 1
cnvMatrix <- cnvMatrix[!is.na(cnvMatrix$Label),]

# Remove "GDC_Aliquot" and "Segment_Mean" columns and change col.names with reordering
cnvMatrix <- cnvMatrix[,-c(1, 6)]
colnames(cnvMatrix) <- c("Chromosome", "Start", "End", "Num.of.Markers", "Sample.Name", "Aberration")
cnvMatrix <- cnvMatrix[, c(5, 1, 2, 3, 4, 6)]

# Substitute Chromosomes "X" and "Y" with "23" and "24"
cnvMatrix[cnvMatrix$Chromosome == "X", "Chromosome"] <- 23
cnvMatrix[cnvMatrix$Chromosome == "Y", "Chromosome"] <- 24
cnvMatrix$Chromosome <- as.integer(cnvMatrix$Chromosome)
write.table(cnvMatrix, file = paste0(cancer, "_cnvMatrix.tsv"), sep = "\t", col.names = TRUE, row.names = FALSE)

# Recurrent CNV identification with GAIA
unique(markersMatrix$Chromosome)
markersMatrix[markersMatrix$Chromosome == "X", "Chromosome"] <- 23
markersMatrix[markersMatrix$Chromosome == "Y", "Chromosome"] <- 24
markersMatrix$Chromosome <- as.integer(markersMatrix$Chromosome)
markersMatrix <- markersMatrix[order(markersMatrix$Start),]
markerID <- paste(markersMatrix$Chromosome, markersMatrix$Start, sep = ":")

# Removed duplicates
markersMatrix <- markersMatrix[!duplicated(markerID),]
rm(markerID)

library(gaia)
set.seed(1)
markers_obj <- load_markers(as.data.frame(markersMatrix))
nbsamples <- length(unique(cnvMatrix$Sample))
cnv_obj <- load_cnv(cnvMatrix, markers_obj, nbsamples)
```

APPENDIX A. (cont.)

```
suppressWarnings(results <- runGAIA(cnv_obj,
                                   markers_obj,
                                   output_file_name = paste0(cancer, "_gaia.txt"),
                                   aberrations = -1, # -1 to all aberrations
                                   chromosomes = -1, # -1 to all chromosomes
                                   approximation = TRUE, # Set to TRUE to speed up the time requirements
                                   num_iterations = 5000, # Reduced to speed up the time requirements
                                   threshold = 0.05)
)
# Set q-value threshold
# You can use a smaller value for your analysis.
threshold <- 0.05

# Plot the results
RecCNV <- t(apply(results,1,as.numeric))
colnames(RecCNV) <- colnames(results)
RecCNV <- cbind(RecCNV, score = 0)
minval <- format(min(RecCNV[RecCNV[, "q-value"] != 0, "q-value"]), scientific = FALSE)
minval <- substring(minval, 1, nchar(minval) - 1)
RecCNV[RecCNV[, "q-value"] == 0, "q-value"] <- as.numeric(minval)
RecCNV[, "score"] <- sapply(RecCNV[, "q-value"], function(x) -log10(as.numeric(x)))
RecCNV[RecCNV[, "q-value"] == as.numeric(minval),]

tiff(file = paste0(cancer, "_CNVplot.tif"), width = 7.5, height = 5.3, units = 'in', res = 300, compression
     = 'lzw')
gaiaCNVplot(RecCNV, threshold)
dev.off()

save(results, RecCNV, threshold, file = paste0(cancer, "_CNV_results.rda"))

## Gene annotation of recurrent CNV

library(GenomicRanges)

# Get gene information from GENCODE using biomaRt
genes <- TCGAbiolinks:::get.GRCh.bioMart(genome = "hg38")
genes <- genes[genes$external_gene_name != "" & genes$chromosome_name %in% c(1:22, "X", "Y"),]
genes[genes$chromosome_name == "X", "chromosome_name"] <- 23
```

APPENDIX A. (cont.)

```
genes[genes$chromosome_name == "Y", "chromosome_name"] <- 24
genes$chromosome_name <- sapply(genes$chromosome_name, as.integer)
genes <- genes[order(genes$start_position),]
genes <- genes[order(genes$chromosome_name),]
genes <- genes[,c("external_gene_name", "chromosome_name", "start_position", "end_position")]
colnames(genes) <- c("GeneSymbol", "Chr", "Start", "End")
genes_GR <- makeGRangesFromDataFrame(genes, keep.extra.columns = TRUE)
save(genes_GR, genes, file = "genes_GR.rda", compress = "xz")

## Recurrent CNV annotation ##

# Get gene information from GENCODE using biomaRt data(genes_GR)

#load(paste0(cancer, "_CNV_results.rda"))
sCNV <- RecCNV[RecCNV[, "q-value"] <= threshold, c(1:4, 6)]
sCNV <- sCNV[order(sCNV[, 3]),]
sCNV <- sCNV[order(sCNV[, 1]),]
colnames(sCNV) <- c("Chr", "Aberration", "Start", "End", "q-value")
sCNV_GR <- makeGRangesFromDataFrame(sCNV, keep.extra.columns = TRUE)

hits <- findOverlaps(genes_GR, sCNV_GR, type = "within")
sCNV_ann <- cbind(sCNV[subjectHits(hits),], genes[queryHits(hits),])
AberrantRegion <- paste0(sCNV_ann[, 1], ":", sCNV_ann[, 3], "-", sCNV_ann[, 4])
GeneRegion <- paste0(sCNV_ann[, 7], ":", sCNV_ann[, 8], "-", sCNV_ann[, 9])
AmpDel_genes <- cbind(sCNV_ann[, c(6, 2, 5)], AberrantRegion, GeneRegion)
AmpDel_genes[AmpDel_genes[, 2] == 0, 2] <- "Del"
AmpDel_genes[AmpDel_genes[, 2] == 1, 2] <- "Amp"
rownames(AmpDel_genes) <- NULL

write.table(AmpDel_genes, file = paste0(cancer, "_AmpDel_genes.tsv"), sep = "\t", col.names = TRUE,
row.names = TRUE)
save(RecCNV, AmpDel_genes, file = paste0(cancer, "_CNV_results_genes.rda"))
```

APPENDIX A. (cont.)

```
## GENE EXPRESSION DATA

library(TCGAbiolinks)
library(SummarizedExperiment)
library(dplyr)
library(DT)

cancer <- "LUAD"

common_patients <- read.csv(file = paste0(cancer, "_common_patients.tsv"), header = FALSE, sep =
"\t")
common_patients <- as.character(unlist(common_patients))

query_exp <- GDCquery(project = paste0("TCGA-", cancer),
                      data.category = "Transcriptome Profiling",
                      data.type = "Gene Expression Quantification",
                      workflow.type = "HTSeq - Counts",
                      barcode = common_patients)

cases_exp <- getResults(query_exp, cols="cases")
cases_exp <- TCGAquery_SampleTypes(cases_exp, c("NT", "TP"))
write.table(cases_exp, file = paste0(cancer, "_cases_exp.tsv"), row.names=FALSE, na="NA", col.names=FALSE, sep="\t")

# Filter TCGA Replicate Samples (cases_exp -> uniq_tsb_exp)
source("tcga_replicateFilter_exp.R")

query_exp <- GDCquery(project = paste0("TCGA-", cancer),
                      data.category = "Transcriptome Profiling",
                      data.type = "Gene Expression Quantification",
                      workflow.type = "HTSeq - Counts",
                      barcode = uniq_tsb_exp)

GDCdownload(query_exp, files.per.chunk = 100)

GeneExp <- GDCprepare(query_exp)
save(GeneExp, file = paste0(cancer, "_GeneExp.rda"))
```

APPENDIX A. (cont.)

```
GeneExp_Matrix <- assay(GeneExp)
save(GeneExp_Matrix, file = paste0(cancer, "_GeneExp_Matrix.rda"))
write.table(GeneExp_Matrix, file= paste0(cancer, "_GeneExp_Matrix.tsv"),
            row.names=TRUE, na="NA",col.names=TRUE, sep="\t")

GeneExp_dataframe <- as.data.frame(colData(GeneExp))
write.table(GeneExp_dataframe, file = paste0(cancer, "_GeneExp_dataframe.tsv"),
            row.names=FALSE, na="NA",col.names=TRUE, sep="\t")

write.table(GeneExp_dataframe[,2:4], file = paste0(cancer, "_targets.tsv"),
            row.names=FALSE, na="NA",col.names=TRUE, sep="\t")

### Differentially Expression Analysis (DEA) by limma-voom

library(limma)
library(edgeR)

targets <- read.delim(paste0(cancer, "_targets.tsv"), sep = "\t")

dge <- DGEList(counts=GeneExp_Matrix)

#filtering to remove rows that consistently have zero or very low counts
keep <- filterByExpr(dge)
dge <- dge[keep,]

#TMM normalization
dgen <- calcNormFactors(dge, method="TMM")
save(dgen, file = paste0(cancer, "_dge_normalized.rda"))
write.table(dgen$counts, file = paste0(cancer, "_dge_normalized.tsv"),
            row.names=TRUE, na="",col.names=TRUE, sep="\t")

design <- model.matrix(~ targets$shortLetterCode)

v <- voom(dgen, design, plot=TRUE)

cor <- duplicateCorrelation(v, design, block = targets$patient)
cor$consensus

v <- voom(dgen, design, block = targets$patient, correlation = cor$consensus)
```

APPENDIX A. (cont.)

```
fit <- lmFit(v, design, block = targets$patient, correlation = cor$consensus)
fit <- eBayes(fit)

DE_genes <- topTable(fit, coef=ncol(design), number=Inf, sort.by="P")
sign_genes <- topTable(fit, coef=ncol(design), number=Inf, sort.by="P", p.value=0.05, lfc=1)
rownames(sign_genes) <- NULL
colnames(sign_genes)[colnames(sign_genes)=="ID"] <- "ensembl_gene_id"
save(sign_genes, file = paste0(cancer, "_sign_genes.rda"))
write.table(sign_genes, file = paste0(cancer, "_sign_genes.tsv"), row.names=FALSE, na="NA", col.names=TRUE, sep="\t")

## Volcano Plot

TCGAVisualize_volcano(DE_genes$logFC, DE_genes$adj.P.Val,
  filename = paste0(cancer, "_VolcanoPlot.png"),
  x.cut = 1,
  y.cut = 0.05,
  color = c("black", "red", "darkgreen"),
  names.size = 2,
  xlab = "Gene expression fold change (Log2)",
  legend = "State",
  title = "Volcano Plot",
  width = 7.5,
  height = 5.3,
  dpi = 300)

##ID conversion
library(biomaRt)
sign_genes_EnsID <- sign_genes$ensembl_gene_id
human_ensembl <- useMart(biomart = "ensembl", dataset = "hsapiens_gene_ensembl")
sign_genes_IDs <- getBM(attributes = c('ensembl_gene_id', 'entrezgene', 'hgnc_symbol'),
  filters = 'ensembl_gene_id',
  values = sign_genes_EnsID,
  mart = human_ensembl)

write.table(sign_genes_IDs, file = paste0(cancer, "_sign_genes_IDs.tsv"),
  row.names=FALSE, col.names=TRUE, na="NA", sep="\t")

sign_genes_2 <- inner_join(sign_genes, sign_genes_IDs, by = "ensembl_gene_id")
```

APPENDIX A. (cont.)

```
sign_genes_2 <- subset(sign_genes_2, select=c("ensembl_gene_id", "entrezgene", "hgnc_symbol", "logFC", "adj.P.Val"))
sign_genes_2 <- sign_genes_2 %>% mutate(ensembl_gene_id = as.character(ensembl_gene_id))
save(sign_genes_2, file = paste0(cancer, "_sign_genes_2.rda"))
write.table(sign_genes_2, file = paste0(cancer, "_sign_genes_2.tsv"),
  row.names=FALSE, na="NA", col.names=TRUE, sep="\t")

sign_down <- sign_genes_2[(sign_genes_2$adj.P.Val<0.05 & sign_genes_2$logFC<(-1)),]
sign_down_head <- head(sign_down, n=10)
write.table(sign_down_head, file = paste0(cancer, "_sign_down_head.tsv"), row.names=FALSE, na="NA", col.names=TRUE, sep="\t")
sign_up <- sign_genes_2[(sign_genes_2$adj.P.Val<0.05 & sign_genes_2$logFC>1),]
sign_up_head <- head(sign_up, n=10)
write.table(sign_up_head, file = paste0(cancer, "_sign_up_head.tsv"), row.names=FALSE, na="NA", col.names=TRUE, sep="\t")

## Enrichment Analysis

EA <- TCGAanalyze_EAcomplete(TFname="Differentially Expressed Genes in Tumor vs Normal",
  RegulonList = sign_genes_2$hgnc_symbol)

tiff(file = paste0(cancer, "_EABarplot_Total%02d.tif"), width = 7.5, height = 5.3, units = 'in', res = 300,
  compression = 'lzw')
TCGAvisualize_EABarplot(tf = rownames(EA$ResBP),
  GOBPTab = EA$ResBP,
  GOCCTab = EA$ResCC,
  GOMFTab = EA$ResMF,
  PathTab = EA$ResPat,
  nRGTab = sign_genes_2$hgnc_symbol,
  nBar = 20,
  text.size = 0.6,
  mfrow = c(1, 1),
  filename = NULL)
dev.off()

EA_up <- TCGAanalyze_EAcomplete(TFname="Up-Regulated Genes in Tumor vs Normal",
  RegulonList = sign_up$hgnc_symbol)
```

APPENDIX A. (cont.)

```
tiff(file = paste0(cancer, "_EAbarplot_Up%02d.tif"), width = 7.5, height = 5.3, units = 'in', res = 300, compression = 'lzw')
```

```
TCGAvisualize_EAbarplot(tf = rownames(EA_up$ResBP),  
  GOBPTab = EA_up$ResBP,  
  GOCCTab = EA_up$ResCC,  
  GOMFTab = EA_up$ResMF,  
  PathTab = EA_up$ResPat,  
  nRGTab = sign_up$hgnc_symbol,  
  nBar = 20,  
  text.size = 0.6,  
  mfrow = c(1, 1),  
  filename = NULL)
```

```
dev.off()
```

```
EA_down <- TCGAanalyze_EAcomplete(TFname="Down-Regulated Genes in Tumor vs Normal",  
  RegulonList = sign_down$hgnc_symbol)
```

```
tiff(file = paste0(cancer, "_EAbarplot_Down%02d.tif"), width = 7.5, height = 5.3, units = 'in', res = 300, compression = 'lzw')
```

```
TCGAvisualize_EAbarplot(tf = rownames(EA_down$ResBP),  
  GOBPTab = EA_down$ResBP,  
  GOCCTab = EA_down$ResCC,  
  GOMFTab = EA_down$ResMF,  
  PathTab = EA_down$ResPat,  
  nRGTab = sign_down$hgnc_symbol,  
  nBar = 20,  
  text.size = 0.6,  
  mfrow = c(1, 1),  
  filename = NULL)
```

```
dev.off()
```

```
## Subnetwork Analysis
```

```
library("DEsubs")
```

```
myRankedList <- setNames(sign_genes_2[, 'adj.P.Val'], sign_genes_2$senseembl_gene_id)
```

APPENDIX A. (cont.)

```
DEsubs.out <- DEsubs(org='hsa',
  mRNAexpr=GeneExp_Matrix,
  mRNAomenclature='ensembl_gene_id',
  pathways='All',
  DEtool=NULL,
  DEpar=0.05,
  CORtool='pearson',
  CORpar=0.6,
  subpathwayType='community.walktrap',
  rankedList=myRankedList,
  verbose=TRUE)

tiff(file = paste0(cancer, "_geneVisualization%02d.tif"), width = 7.5, height = 5.3, units = 'in', res = 300, compression = 'lzw')
suppressWarnings(geneVisualization(DEsubs.out=DEsubs.out, top=10,
  measures.topological=c( 'degree', 'betweenness', 'closeness',
    'eccentricity', 'page_rank', 'hub_score'),
  measures.functional='all',
  size.topological=c(5,4),
  size.functional=c(7,4),
  size.barplot=c(5,6),
  export='plot',
  verbose=FALSE)
)
dev.off()

DEsubs_genes.df <- do.call("rbind", lapply(DEsubs.out[["subAnalysis.community.walktrap"]], as.data.frame))
write.table(DEsubs_genes.df, file="DEsubs_genes.tsv", na="NA", col.names=FALSE, sep="\t")

tiff(file = paste0(cancer, "_sub%03d.tif"), width = 7.5, height = 5.3, units = 'in', res = 300, compression = 'lzw')
for (i in 1:length(DEsubs.out[["subAnalysis.community.walktrap"]])) {
  subpathwayToGraph(DEsubs.out=DEsubs.out,
    submethod='community.walktrap',
    subname=paste0('sub', i), verbose=FALSE,
    export='plot')
}
dev.off()
```

APPENDIX A. (cont.)

```
tiff(file = paste0(cancer, "_organismVisualization%02d.tif"), width = 7.5, height = 5.3, units = 'in', res
= 300, compression = 'lzw')
organismVisualization(DEsubs.out=DEsubs.out,
  references=c('KEGG', 'TF', 'Drug_DrugBank'),
  topSubs=10,
  topTerms=20,
  export='plot',
  verbose=FALSE)
dev.off()

# Cox Proportional Hazards Model -----

library(dplyr)
library(TCGAbiolinks)
library(survival)
library(survminer)
library(maftools)

cancer <- "LUAD"

common_patients <- read.csv(file = paste0(cancer, "_common_patients.tsv"), header = FALSE, sep =
"\t", stringsAsFactors=FALSE)
common_patients <- as.character(unlist(common_patients))

clinic <- GDCquery_clinic(project = paste0("TCGA-", cancer), type = "clinical")

cfu <- clinic[clinic[, "bcr_patient_barcode"] %in% substr(common_patients, 1, 12), ]

cfu <- as.data.frame(subset(cfu, select = c("bcr_patient_barcode", "days_to_last_follow_up",
  "days_to_death", "vital_status",
  "age_at_diagnosis", "gender",
  "tumor_stage", "years_smoked")))
rownames(cfu) <- cfu$bcr_patient_barcode
cfu <- cfu[, -1]
colnames(cfu) <- c("time", "timeDead", "status", "age", "gender", "tumor_stage", "years_smoked")
cfu[which(cfu$status == "dead"), "time"] <- cfu[which(cfu$status == "dead"), "timeDead"]
cfu[which(cfu$status == "alive"), "status"] <- 0
cfu[which(cfu$status == "dead"), "status"] <- 1
cfu$gender <- tolower(cfu$gender)
```

APPENDIX A. (cont.)

```
cfu <- cfu[, -2]
cfu$time <- as.numeric(cfu$time)
cfu$status <- as.numeric(cfu$status)
cfu$age <- as.numeric(cfu$age)
cfu[cfu == "stage i"] <- 1
cfu[cfu == "stage ia"] <- 1
cfu[cfu == "stage ib"] <- 1
cfu[cfu == "stage ii"] <- 2
cfu[cfu == "stage iia"] <- 2
cfu[cfu == "stage iib"] <- 2
cfu[cfu == "stage iii"] <- 3
cfu[cfu == "stage iiia"] <- 3
cfu[cfu == "stage iiib"] <- 3
cfu[cfu == "stage iv"] <- 4
cfu[cfu == "not reported"] <- NA

cfu2 <- as.data.frame(subset(cfu, select = c("time", "status")))

## Univariate Cox regression of clinic variables
fit1 <- coxph(formula = Surv(time, status) ~ age, data = cfu)
summary(fit1)

fit2 <- coxph(formula = Surv(time, status) ~ gender, data = cfu)
summary(fit2)

fit3 <- coxph(formula = Surv(time, status) ~ tumor_stage, data = cfu)
summary(fit3)

fit4 <- coxph(formula = Surv(time, status) ~ years_smoked, data = cfu)
summary(fit4)

fit <- survfit (Surv(time, status) ~ tumor_stage, data = cfu)
tiff(file = paste0(cancer, "_survival_tumor_stage.tif"), width = 7.5, height = 5.3, units = 'in', res = 300,
compression = 'lzw')
ggsurvplot(fit, data = cfu, pval = TRUE, legend.labs=c("Stage I", "Stage II", "Stage III", "Stage IV"),
ggtheme = theme_survminer())
dev.off()
```

APPENDIX A. (cont.)

```
## Univariate Cox regression of CNVs

cnvMatrix <- read.csv(file = paste0(cancer, "_cnvMatrix.tsv"), sep = "\t", stringsAsFactors=FALSE)
cnvMatrix <- cnvMatrix[order(cnvMatrix$Chromosome, cnvMatrix$Start),]
cnv_results <- read.csv(file = paste0(cancer, "_AmpDel_genes.tsv"), sep = "\t", stringsAsFactors=FALSE)
cnv_results <- cnv_results[!duplicated(cnv_results$GeneSymbol),]
nrow <- length(common_patients)
ncol <- length(unique(cnv_results$GeneSymbol))
cnv_gene_matrix <- data.frame(matrix(data="A", ncol = ncol, nrow = nrow), stringsAsFactors=FALSE)
colnames(cnv_gene_matrix) <- unique(cnv_results$GeneSymbol)
rownames(cnv_gene_matrix) <- common_patients

for (i in 1:nrow(cnv_results)){
  string <- cnv_results$GeneRegion[i]
  pattern <- "[[:punct:]]"
  splits <- strsplit(string, pattern)
  Chr <- splits[[1]][1]
  Chr_start <- splits[[1]][2]
  Chr_end <- splits[[1]][3]
  gene_name <- cnv_results$GeneSymbol[i]
  aberration <- cnv_results$Aberration[i]
  for (j in 1:nrow(cnvMatrix)){
    if (as.numeric(Chr) == cnvMatrix$Chromosome[j] && as.numeric(Chr_start) >= cnvMatrix$Start[j] && as.numeric(Chr_end) <= cnvMatrix$End[j]){
      patient <- substr(cnvMatrix$Sample.Name[j], 1, 12)
      if(cnvMatrix$Aberration[j] == 0){
        gene_abr <- "Del"
      } else {
        gene_abr <- "Amp"
      }
      if(gene_abr == aberration){
        cnv_gene_matrix[patient, gene_name] <- "C"
      }
    }
  }
}
```

APPENDIX A. (cont.)

```
res_cnv <- data.frame(matrix(ncol = 6, nrow = ncol(cnv_gene_matrix)), stringsAsFactors=FALSE)
colnames(res_cnv) <- c("gene.name", "aberration", "beta", "HR (95% CI for HR)", "wald.test", "p.value")

for (k in 1:ncol(cnv_gene_matrix)) {
  data <- cbind(cfu2, gene=cnv_gene_matrix[,k])
  data <- data[which(data$gene != "-Inf"), ]
  fit <- coxph(formula = Surv(time, status) ~ gene, data = data)
  cox <- summary(fit)
  gene_name <- colnames(cnv_gene_matrix[k])
  abr <- cnv_results[which(cnv_results$GeneSymbol == gene_name), "Aberration"]
  p.value <- as.numeric(signif(cox$wald["pvalue"], digits=2))
  wald.test <- as.numeric(signif(cox$wald["test"], digits=2))
  beta <- signif(cox$coef[1], digits=2)
  HR <- signif(cox$coef[2], digits=2)
  HR.confint.lower <- signif(cox$conf.int["lower .95"], 2)
  HR.confint.upper <- signif(cox$conf.int["upper .95"], 2)
  HR <- paste0(HR, " (", HR.confint.lower, "-", HR.confint.upper, ")")
  res_cnv[k,] <- c(gene_name, abr, beta, HR, wald.test, p.value)
}

res_cnv_sign <- res_cnv[which(res_cnv$p.value <= 0.05),]
res_cnv_sign <- res_cnv_sign[order(res_cnv_sign$p.value),]
rownames(res_cnv_sign) <- NULL

## Univariate Cox regression of SNVs

load(paste0(cancer, "_maf.RData"))

somniaclust_results <- read.csv(file = paste0(cancer, "_SomInaClust_results_summary_table.txt"), sep
="\\t", stringsAsFactors=FALSE)
somniaclust_genes <- sominaclust_results$Gene

snv_matrix <- data.frame(matrix(data="A", ncol = length(somniaclust_genes), nrow = length(comm
on_patients)), stringsAsFactors=FALSE)
colnames(snv_matrix) <- sominaclust_genes
rownames(snv_matrix) <- common_patients
```

APPENDIX A. (cont.)

```
mafselected <- subsetMaf(maf = maf, genes = sominaclust_genes,
                        mafObj = FALSE, isTCGA = TRUE,
                        query = "Variant_Classification != 'Silent'")

for (l in 1:nrow(mafselected)) {
  gene <- mafselected$Hugo_Symbol[l]
  patient <- substr(mafselected$Tumor_Sample_Barcode[l],1,12)
  snv_matrix[patient,gene] <- "M" # for mutation
}

res_snv <- data.frame(matrix(ncol = 5, nrow = ncol(snv_matrix)), stringsAsFactors=FALSE)
colnames(res_snv) <- c("gene.name", "beta", "HR (95% CI for HR)", "wald.test", "p.value")

for (k in 1:ncol(snv_matrix)) {
  data <- cbind(cfu2, gene=snv_matrix[,k])
  data <- data[which(data$gene != "-Inf"), ]
  fit <- coxph(formula = Surv(time, status) ~ gene, data = data)
  cox <- summary(fit)
  gene_name <- colnames(snv_matrix[k])
  p.value <- as.numeric(signif(cox$wald["pvalue"], digits=2))
  wald.test <- as.numeric(signif(cox$wald["test"], digits=2))
  beta <- signif(cox$coef[1], digits=2)
  HR <- signif(cox$coef[2], digits=2)
  HR.confint.lower <- signif(cox$conf.int["lower .95"], 2)
  HR.confint.upper <- signif(cox$conf.int["upper .95"], 2)
  HR <- paste0(HR, " (",
               HR.confint.lower, "-", HR.confint.upper, ")")
  res_snv[k,] <- c(gene_name, beta, HR, wald.test, p.value)
}

res_snv_sign <- res_snv[which(res_snv$p.value <= 0.05),]
res_snv_sign <- res_snv_sign[order(res_snv_sign$p.value),]
rownames(res_snv_sign) <- NULL
```

APPENDIX A. (cont.)

```
## Univariate Cox regression of DEGs

load(paste0(cancer, "_dge_normalized.rda"))

degs <- read.csv(file = paste0(cancer, "_sign_genes_2.tsv"), sep = "\t", na.strings=c("", "NA"), stringsAsFactors=FALSE)
degs <- na.omit(degs)
degs_genes <- degs$hgnc_symbol

degs_matrix <- data.frame(matrix(data=0, ncol = length(degs_genes), nrow = length(common_patients)), stringsAsFactors=FALSE)
colnames(degs_matrix) <- degs_genes
rownames(degs_matrix) <- common_patients

cont_samples <- TCGAquery_SampleTypes(colnames(dgen$counts), typesample = c("NT"))
tumor_samples <- TCGAquery_SampleTypes(colnames(dgen$counts), typesample = c("TP"))

for (i in 1:nrow(degs)) {
  gene <- degs$hgnc_symbol[i]
  ensmb1 <- degs$ensembl_gene_id[i]
  for (j in 1:length(tumor_samples)) {
    sample <- tumor_samples[j]
    patient <- substr(sample,1,12)
    degs_matrix[patient, gene] <- dgen$counts[ensmb1, sample]
  }
}

degs_matrix <- log2(degs_matrix+1)
res_degs <- data.frame(matrix(ncol = 5, nrow = ncol(degs_matrix)), stringsAsFactors=FALSE)
colnames(res_degs) <- c("gene.name", "beta", "HR (95% CI for HR)", "wald.test", "p.value")

for (i in 1:ncol(degs_matrix)) {
  data <- cbind(cfu2, gene=degs_matrix[,i])
  data <- data[which(data$gene != "-Inf"), ]
  fit <- coxph(formula = Surv(time, status) ~ gene, data = data)
  cox <- summary(fit)
  gene_name <- colnames(degs_matrix[i])
  p.value <- as.numeric(signif(cox$wald["pvalue"], digits=2))
  wald.test <- as.numeric(signif(cox$wald["test"], digits=2))
  beta <- signif(cox$coef[1], digits=2)
```

APPENDIX A. (cont.)

```
HR <- signif(cox$coef[2], digits=2)
HR.confint.lower <- signif(cox$conf.int[, "lower .95"], 2)
HR.confint.upper <- signif(cox$conf.int[, "upper .95"], 2)
HR <- paste0(HR, " (",
              HR.confint.lower, "-", HR.confint.upper, ")")
res_degs[i,] <- c(gene_name, beta, HR, wald.test, p.value)
}

res_degs_sign <- res_degs[which(res_degs$p.value <= 0.05),]
res_degs_sign <- res_degs_sign[order(res_degs_sign$p.value),]
rownames(res_degs_sign) <- NULL

## Univariate Cox regression of DEsubs genes
desubs <- read.csv("LUAD_DEsubs_genes.tsv", sep = "\t", col.names = c("sub", "entrez"))

desubs_IDs <- merge(desubs, degs, by.x = "entrez", by.y = "entrezgene")
desubs_IDs <- desubs_IDs[-c(2,5,6)]

degs_matrix2 <- data.frame(matrix(data=0, ncol = nrow(desubs_IDs), nrow = length(common_patients)), stringsAsFactors=FALSE)
colnames(degs_matrix2) <- desubs_IDs$hgnc_symbol
rownames(degs_matrix2) <- common_patients

cont_samples <- TCGAquery_SampleTypes(colnames(dgen$counts), typesample = c("NT"))
tumor_samples <- TCGAquery_SampleTypes(colnames(dgen$counts), typesample = c("TP"))

for (i in 1:nrow(desubs_IDs)) {
  hgnc <- desubs_IDs$hgnc_symbol[i]
  ensmb1 <- desubs_IDs$ensembl_gene_id[i]
  for (j in 1:length(tumor_samples)) {
    sample <- tumor_samples[j]
    patient <- substr(sample,1,12)
    degs_matrix2[patient,hgnc] <- dgen$counts[ensmb1,sample]
  }
}

degs_matrix2 <- log2(degs_matrix2+1)
res_degs2 <- data.frame(matrix(ncol = 5, nrow = ncol(degs_matrix2)), stringsAsFactors=FALSE)
colnames(res_degs2) <- c("gene.name", "beta", "HR (95% CI for HR)", "wald.test", "p.value")
```

APPENDIX A. (cont.)

```
for (i in 1:ncol(degs_matrix2)) {
  data <- cbind(cfu2, gene=degs_matrix2[,i])
  data <- data[which(data$gene != "-Inf"), ]
  fit <- coxph(formula = Surv(time, status) ~ gene, data = data)
  cox <- summary(fit)
  gene_name <- colnames(degs_matrix2[i])
  p.value <- as.numeric(signif(cox$wald["pvalue"], digits=2))
  wald.test <- as.numeric(signif(cox$wald["test"], digits=2))
  beta <- signif(cox$coef[1], digits=2)
  HR <- signif(cox$coef[2], digits=2)
  HR.confint.lower <- signif(cox$conf.int["lower .95"], 2)
  HR.confint.upper <- signif(cox$conf.int["upper .95"], 2)
  HR <- paste0(HR, " (",
    HR.confint.lower, "-", HR.confint.upper, ")")
  res_degs2[i,] <- c(gene_name, beta, HR, wald.test, p.value)
}

res_degs_sign2 <- res_degs2[which(res_degs2$p.value <= 0.05),]
res_degs_sign2 <- res_degs_sign2[order(res_degs_sign2$p.value),]
rownames(res_degs_sign2) <- NULL

## Multivariate Cox regression analysis -----

res_cnv_sign_matrix <- cnv_gene_matrix[,res_cnv_sign$gene.name]
c <- colnames(res_cnv_sign_matrix)
c <- gsub("[ -]", "_", c)
colnames(res_cnv_sign_matrix) <- c

res_degs_sign_matrix <- degs_matrix2[,res_degs_sign2$gene.name]
c2 <- colnames(res_degs_sign_matrix)
c2 <- gsub("[ -]", "_", c2)
colnames(res_degs_sign_matrix) <- c2

mv_matrix <- cbind(snv_matrix, res_degs_sign_matrix, res_cnv_sign_matrix)
df <- cbind(cfu2, mv_matrix)
f <- as.formula(paste("Surv(time, status) ~ ", paste(colnames(mv_matrix), collapse= "+")))
fit_mv <- coxph(f, data=df)
mv_res.cox <- summary(fit_mv)
mv_res.cox$coefficients
```

APPENDIX A. (cont.)

```
mv_cox_table <- mv_res.cox$coefficients[,c(2,5)]
mv_cox_table <- mv_cox_table[which(mv_cox_table[,2] <= 0.05),]
mv_cox_table <- mv_cox_table[order(mv_cox_table[,2]),]
colnames(mv_cox_table) <- c("HR", "p.value")
mv_cox_table
mv_res.cox$waldtest
write.table(mv_cox_table, file = paste0(cancer, "_mv_cox_table.tsv"), row.names=TRUE, na="NA", col.names=TRUE, sep="\t")

df1 <- cbind(cfu2, snv_matrix)
f1 <- as.formula(paste("Surv(time, status) ~ ", paste(colnames(snv_matrix), collapse="+")))
fit_mv1 <- coxph(f1, data=df1)
mv_res.cox1 <- summary(fit_mv1)
mv_res.cox1$coefficients
mv_cox1_table <- mv_res.cox1$coefficients[,c(2,5)]
mv_cox1_table <- mv_cox1_table[which(mv_cox1_table[,2] <= 0.05),]
mv_cox1_table <- mv_cox1_table[order(mv_cox1_table[,2]),]
colnames(mv_cox1_table) <- c("HR", "p.value")
mv_cox1_table
mv_res.cox1$waldtest

df2 <- cbind(cfu2, res_cnv_sign_matrix)
f2 <- as.formula(paste("Surv(time, status) ~ ", paste(colnames(res_cnv_sign_matrix), collapse="+")))
fit_mv2 <- coxph(f2, data=df2)
mv_res.cox2 <- summary(fit_mv2)
mv_res.cox2$coefficients
mv_cox2_table <- mv_res.cox2$coefficients[,c(2,5)]
mv_cox2_table <- mv_cox2_table[which(mv_cox2_table[,2] <= 0.05),]
mv_cox2_table <- mv_cox2_table[order(mv_cox2_table[,2]),]
colnames(mv_cox2_table) <- c("HR", "p.value")
mv_cox2_table
mv_res.cox2$waldtest

df3 <- cbind(cfu2, res_degs_sign_matrix)
f3 <- as.formula(paste("Surv(time, status) ~ ", paste(colnames(res_degs_sign_matrix), collapse="+")))
fit_mv3 <- coxph(f3, data=df3)
mv_res.cox3 <- summary(fit_mv3)
```

APPENDIX A. (cont.)

```
mv_res.cox3$coefficients
mv_cox3_table <- mv_res.cox3$coefficients[,c(2,5)]
mv_cox3_table <- mv_cox3_table[which(mv_cox3_table[,2] <= 0.05),]
mv_cox3_table <- mv_cox3_table[order(mv_cox3_table[,2]),]
colnames(mv_cox3_table) <- c("HR", "p.value")
mv_cox3_table
mv_res.cox3$waldtest

# Hierarchical Clustering of patients and Survival Analysis -----

# HeatMap and Survival before CPHM

# Limma DEGs
sig_genes_exp_tumor <- degs_matrix
sig_genes_exp_tumor <- mapply(sig_genes_exp_tumor, FUN=as.numeric)
rownames(sig_genes_exp_tumor) <- rownames(degs_matrix)

tiff(file = paste0(cancer, "_Heatmap_total_BC.tif"), width = 7.5, height = 5.3, units = 'in', res = 300, co
mpression = 'lzw')
heatmap(sig_genes_exp_tumor, scale = "column", margins = c(5.5,7),
        xlab = "Genes", ylab = "Patients")
dev.off()

hc1 <- TCGAanalyze_Clustering(tabDF = sig_genes_exp_tumor,
                             method = "hclust",
                             methodHC = "ward.D2")

tiff(file = paste0(cancer, "_HC_total_BC.tif"), width = 7.5, height = 5.3, units = 'in', res = 300, compres
sion = 'lzw')
plot(hc1, cex=0.5)
dev.off()

hclusters1 = cutree(hc1, k=2)
cfu3 <- cbind(cfu2, HCgroup=hclusters1)

tiff(file = paste0(cancer, "_SV_total_BC.tif"), width = 7.5, height = 5.3, units = 'in', res = 300, compres
sion = 'lzw')
ggsurvplot(survfit(Surv(time, status) ~ HCgroup, data = cfu3), data = cfu3, pval = TRUE)
dev.off()
```

APPENDIX A. (cont.)

```
#DEsubs DEGs
sub_genes_exp_tumor <- degs_matrix2
sub_genes_exp_tumor <- mapply(sub_genes_exp_tumor, FUN=as.numeric)
rownames(sub_genes_exp_tumor) <- rownames(degs_matrix2)

tiff(file = paste0(cancer, "_Heatmap_sub_BC.tif"), width = 7.5, height = 5.3, units = 'in', res = 300, co
mpression = 'lzw')
heatmap(sub_genes_exp_tumor, scale = "column", margins = c(5.5,7),
  xlab = "Genes", ylab = "Patients")
dev.off()

hc2 <- TCGAanalyze_Clustering(tabDF = sub_genes_exp_tumor,
  method = "hclust",
  methodHC = "ward.D2")

tiff(file = paste0(cancer, "_HC_sub_BC.tif"), width = 7.5, height = 5.3, units = 'in', res = 300, compres
sion = 'lzw')
plot(hc2,cex=0.5)
dev.off()

hclusters2 = cutree(hc2, k=2)
cfu4 <- cbind(cfu2, HCgroup=hclusters2)

tiff(file = paste0(cancer, "_SV_sub_BC.tif"), width = 7.5, height = 5.3, units = 'in', res = 300, compress
ion = 'lzw')
ggsurvplot(survfit(Surv(time, status) ~ HCgroup, data = cfu4), data = cfu4, pval = TRUE)
dev.off()

# HeatMap and Survival after CPHM
#DEsubs DEGs mv alone
sub_genes_exp_cox <- degs_matrix2[,colnames(degs_matrix2) %in% rownames(mv_cox3_table)]
sub_genes_exp_cox <- mapply(sub_genes_exp_cox, FUN=as.numeric)
rownames(sub_genes_exp_cox) <- rownames(degs_matrix2)

tiff(file = paste0(cancer, "_Heatmap_sub_AC.tif"), width = 7.5, height = 5.3, units = 'in', res = 300, co
mpression = 'lzw')
heatmap(sub_genes_exp_cox, scale = "column", margins = c(5.5,7),
  xlab = "Genes", ylab = "Patients")
dev.off()
```

APPENDIX A. (cont.)

```
hc3 <- TCGAanalyze_Clustering(tabDF = sub_genes_exp_cox,  
                              method = "hclust",  
                              methodHC = "ward.D2")  
  
tiff(file = paste0(cancer, "_HC_sub_AC.tif"), width = 7.5, height = 5.3, units = 'in', res = 300, compression = 'lzw')  
plot(hc3,cex=0.5)  
dev.off()  
  
hclusters3 = cutree(hc3, k=2)  
cfu5 <- cbind(cfu2, HCgroup=hclusters3)  
  
tiff(file = paste0(cancer, "_SV_sub_AC.tif"), width = 7.5, height = 5.3, units = 'in', res = 300, compression = 'lzw')  
ggsurvplot(survfit(Surv(time, status) ~ HCgroup, data = cfu5), data = cfu5, pval = TRUE)  
dev.off()  
  
#DEsubs DEGs mv total  
mv_genes_exp <- degs_matrix2[,colnames(degs_matrix2) %in% rownames(mv_cox_table)]  
mv_genes_exp <- mapply(mv_genes_exp, FUN=as.numeric)  
rownames(mv_genes_exp) <- rownames(degs_matrix2)  
  
tiff(file = paste0(cancer, "_Heatmap_mv_AC.tif"), width = 7.5, height = 5.3, units = 'in', res = 300, compression = 'lzw')  
heatmap(mv_genes_exp, scale = "column", margins = c(5.5,7),  
        xlab = "Genes", ylab = "Patients")  
dev.off()  
  
hc4 <- TCGAanalyze_Clustering(tabDF = mv_genes_exp,  
                              method = "hclust",  
                              methodHC = "ward.D2")  
  
tiff(file = paste0(cancer, "_HC_mv_AC.tif"), width = 7.5, height = 5.3, units = 'in', res = 300, compression = 'lzw')  
plot(hc4,cex=0.5)  
dev.off()  
  
hclusters4 = cutree(hc4, k=2)  
cfu6 <- cbind(cfu2, HCgroup=hclusters4)
```

APPENDIX A. (cont.)

```
tiff(file = paste0(cancer, "_SV_mv_AC.tif"), width = 7.5, height = 5.3, units = 'in', res = 300, compress
ion = 'lzw')

ggsurvplot(survfit(Surv(time, status) ~ HCgroup, data = cfu6), data = cfu6, pval = TRUE)
dev.off()

#ENRCHMENT

mv_cox_genes <- rownames(mv_cox_table)
EA_mv <- TCGAanalyze_EAcomplete(TFname="Pathways of candidate driver genes in tumor samples",
                                RegulonList = mv_cox_genes)

tiff(file = paste0(cancer, "_EAbarplot_MV_total%02d.tif"), width = 7.5, height = 5.3, units = 'in', res =
300, compression = 'lzw')
TCGAvisualize_EAbarplot(tf = rownames(EA_mv$ResBP),
                        GOBPTab = EA_mv$ResBP,
                        GOCCTab = EA_mv$ResCC,
                        GOMFTab = EA_mv$ResMF,
                        PathTab = EA_mv$ResPat,
                        nRGTab = mv_cox_genes,
                        nBar = 20,
                        text.size = 0.6,
                        mfrow = c(1, 1),
                        filename = NULL)
dev.off()

library(biomaRt)
human_ensembl <- useMart(biomart = "ensembl", dataset = "hsapiens_gene_ensembl")
mv_cox_genes_IDs <- getBM(attributes = c('entrezgene','hgnc_symbol'),
                          filters = 'hgnc_symbol',
                          values = mv_cox_genes,
                          mart = human_ensembl)

write.table(mv_cox_genes_IDs, file = paste0(cancer, "_mv_cox_genes_IDs.tsv"),
            row.names=FALSE, col.names=TRUE, na="NA", sep="\t")

mv_cox_genes_logfc <- degs[degs[, "hgnc_symbol"] %in% mv_cox_genes, ]
mv_cox_genes_logfc
```

APPENDIX A. (cont.)

Filtering of TCGA Replicate Samples

Filter TCGA Replicate Samples

Author: Shixiang Wang <w_shixiang@163.com>

The filter rule following broad institute says:

In many instances there is more than one aliquot for a given combination of individual, platform, and data type. However, only one aliquot may be ingested into Firehose. Therefore, a set of precedence rules are applied to select the most scientifically advantageous one among them. Two filters are applied to achieve this aim: an Analyte Replicate Filter and a Sort Replicate Filter.

Analyte Replicate Filter

The following precedence rules are applied when the aliquots have differing analytes. For RNA aliquots, T analytes are dropped in preference to H and R analytes, since T is the inferior extraction protocol. If H and R are encountered, H is the chosen analyte. This is somewhat arbitrary and subject to change since it is not clear at present whether H or R is the better protocol. If there are multiple aliquots associated with the chosen RNA analyte, the aliquot with the later plate number is chosen. For DNA aliquots, D analytes (native DNA) are preferred over G, W, or X (whole-genome amplified) analytes, unless the G, W, or X analyte sample has a higher plate number.

Sort Replicate Filter

The following precedence rules are applied when the analyte filter still produces more than one sample. The sort filter chooses the aliquot with the highest lexicographical sort value, to ensure that the barcode with the highest portion and/or plate number is selected when all other barcode fields are identical.

Ref Link: <<https://confluence.broadinstitute.org/display/GDAC/FAQ#FAQ-sampleTypesQWhatTCGAsampletypesareFirehosepipelinesexecutedupon>>

Filter FFPE samples, provided in <http://gdac.broadinstitute.org/runs/sampleReports/latest/FPPP_FFPE_Cases.html>

Please note:

Basically, the user provides tsb and analyte_target is fine. If you want to filter FFPE samples, please set filter_FFPE and full_barcode all to TRUE, and tsb must have nchar of 28

DNA analyte replicate filter

> tcga_replicateFilter(tsb = c("TCGA-55-7913-01B-11D-2237-01", "TCGA-55-7913-01B-11X-2237-01", "TCGA-55-7913-01B-11D-2237-01"))

APPENDIX A. (cont.)

```
# > tcga_replicateFilter(tsb = c("TCGA-55-7913-01B-11T-2237-01", "TCGA-55-7913-01B-11T-2237-01", "TCGA-55-7913-01B-11D-2237-01"), analyte_target = "DNA")

# RNA analyte replicate filter
# > tcga_replicateFilter(tsb = c("TCGA-55-7913-01B-11H-2237-01", "TCGA-55-7913-01B-11R-2237-01", "TCGA-55-7913-01B-11T-2237-01"), analyte_target = "RNA")
# > tcga_replicateFilter(tsb = c("TCGA-55-7913-01B-11R-2237-01", "TCGA-55-7913-01B-11R-2237-01", "TCGA-55-7913-01B-11T-2237-01"), analyte_target = "RNA")
# > tcga_replicateFilter(tsb = c("TCGA-55-7913-01B-11H-2237-01", "TCGA-55-7913-01B-11R-2237-01", "TCGA-55-7913-01B-11T-2237-01", "TCGA-44-2656-01B-06D-A271-08", "TCGA-44-2656-01B-06D-A273-01"), analyte_target = "RNA", filter_FFPE = TRUE, full_barcode = TRUE)

tcga_replicateFilter = function(tsb, analyte_target=c("DNA","RNA"), decreasing=TRUE, analyte_position=20, plate=c(22,25), portion=c(18,19), filter_FFPE=FALSE, full_barcode=FALSE){
# Basically, the user provides tsb and analyte_target is fine. If you want to filter FFPE samples, please set filter_FFPE and full_barcode all to TRUE, and tsb must have nchar of 28

  analyte_target = match.arg(analyte_target)
  # Strings in R are largely lexicographic
  # see ??base::Comparison

  # Filter FFPE samples
  # provide by <http://gdac.broadinstitute.org/runs/sampleReports/latest/FPPP\_FFPE\_Cases.html>
  if(full_barcode & filter_FFPE){
    ffpe = c("TCGA-44-2656-01B-06D-A271-08", "TCGA-44-2656-01B-06D-A273-01",
              "TCGA-44-2656-01B-06D-A276-05", "TCGA-44-2656-01B-06D-A27C-26",
              "TCGA-44-2656-01B-06R-A277-07", "TCGA-44-2662-01B-02D-A271-08",
              "TCGA-44-2662-01B-02D-A273-01", "TCGA-44-2662-01B-02R-A277-07",
              "TCGA-44-2665-01B-06D-A271-08", "TCGA-44-2665-01B-06D-A273-01",
              "TCGA-44-2665-01B-06D-A276-05", "TCGA-44-2665-01B-06R-A277-07",
              "TCGA-44-2666-01B-02D-A271-08", "TCGA-44-2666-01B-02D-A273-01",
              "TCGA-44-2666-01B-02D-A276-05", "TCGA-44-2666-01B-02D-A27C-26",
              "TCGA-44-2666-01B-02R-A277-07", "TCGA-44-2668-01B-02D-A271-08",
              "TCGA-44-2668-01B-02D-A273-01", "TCGA-44-2668-01B-02D-A276-05",
              "TCGA-44-2668-01B-02D-A27C-26", "TCGA-44-2668-01B-02R-A277-07",
              "TCGA-44-3917-01B-02D-A271-08", "TCGA-44-3917-01B-02D-A273-01",
              "TCGA-44-3917-01B-02D-A276-05", "TCGA-44-3917-01B-02D-A27C-26",
              "TCGA-44-3917-01B-02R-A277-07", "TCGA-44-3918-01B-02D-A271-08",
              "TCGA-44-3918-01B-02D-A273-01", "TCGA-44-3918-01B-02D-A276-05",
```

APPENDIX A. (cont.)

"TCGA-44-3918-01B-02D-A27C-26", "TCGA-44-3918-01B-02R-A277-07",
"TCGA-44-4112-01B-06D-A271-08", "TCGA-44-4112-01B-06D-A273-01",
"TCGA-44-4112-01B-06D-A276-05", "TCGA-44-4112-01B-06D-A27C-26",
"TCGA-44-4112-01B-06R-A277-07", "TCGA-44-5645-01B-04D-A271-08",
"TCGA-44-5645-01B-04D-A273-01", "TCGA-44-5645-01B-04D-A276-05",
"TCGA-44-5645-01B-04D-A27C-26", "TCGA-44-5645-01B-04R-A277-07",
"TCGA-44-6146-01B-04D-A271-08", "TCGA-44-6146-01B-04D-A273-01",
"TCGA-44-6146-01B-04D-A276-05", "TCGA-44-6146-01B-04D-A27C-26",
"TCGA-44-6146-01B-04R-A277-07", "TCGA-44-6146-01B-04R-A27D-13",
"TCGA-44-6147-01B-06D-A271-08", "TCGA-44-6147-01B-06D-A273-01",
"TCGA-44-6147-01B-06D-A276-05", "TCGA-44-6147-01B-06D-A27C-26",
"TCGA-44-6147-01B-06R-A277-07", "TCGA-44-6147-01B-06R-A27D-13",
"TCGA-44-6775-01C-02D-A271-08", "TCGA-44-6775-01C-02D-A273-01",
"TCGA-44-6775-01C-02D-A276-05", "TCGA-44-6775-01C-02D-A27C-26",
"TCGA-44-6775-01C-02R-A277-07", "TCGA-44-6775-01C-02R-A27D-13",
"TCGA-A6-2674-01B-04D-A270-10", "TCGA-A6-2674-01B-04R-A277-07",
"TCGA-A6-2677-01B-02D-A270-10", "TCGA-A6-2677-01B-02D-A274-01",
"TCGA-A6-2677-01B-02D-A27A-05", "TCGA-A6-2677-01B-02D-A27E-26",
"TCGA-A6-2677-01B-02R-A277-07", "TCGA-A6-2684-01C-08D-A270-10",
"TCGA-A6-2684-01C-08D-A274-01", "TCGA-A6-2684-01C-08D-A27A-05",
"TCGA-A6-2684-01C-08D-A27E-26", "TCGA-A6-2684-01C-08R-A277-07",
"TCGA-A6-3809-01B-04D-A270-10", "TCGA-A6-3809-01B-04D-A274-01",
"TCGA-A6-3809-01B-04D-A27A-05", "TCGA-A6-3809-01B-04D-A27E-26",
"TCGA-A6-3809-01B-04R-A277-07", "TCGA-A6-3810-01B-04D-A270-10",
"TCGA-A6-3810-01B-04D-A274-01", "TCGA-A6-3810-01B-04D-A27A-05",
"TCGA-A6-3810-01B-04D-A27E-26", "TCGA-A6-3810-01B-04R-A277-07",
"TCGA-A6-5656-01B-02D-A270-10", "TCGA-A6-5656-01B-02D-A274-01",
"TCGA-A6-5656-01B-02D-A27A-05", "TCGA-A6-5656-01B-02D-A27E-26",
"TCGA-A6-5656-01B-02R-A277-07", "TCGA-A6-5656-01B-02R-A27D-13",
"TCGA-A6-5659-01B-04D-A270-10", "TCGA-A6-5659-01B-04D-A274-01",
"TCGA-A6-5659-01B-04D-A27A-05", "TCGA-A6-5659-01B-04D-A27E-26",
"TCGA-A6-5659-01B-04R-A277-07", "TCGA-A6-6650-01B-02D-A270-10",
"TCGA-A6-6650-01B-02D-A274-01", "TCGA-A6-6650-01B-02D-A27A-05",
"TCGA-A6-6650-01B-02D-A27E-26", "TCGA-A6-6650-01B-02R-A277-07",
"TCGA-A6-6650-01B-02R-A27D-13", "TCGA-A6-6780-01B-04D-A270-10",
"TCGA-A6-6780-01B-04D-A274-01", "TCGA-A6-6780-01B-04D-A27A-05",
"TCGA-A6-6780-01B-04D-A27E-26", "TCGA-A6-6780-01B-04R-A277-07",
"TCGA-A6-6780-01B-04R-A27D-13", "TCGA-A6-6781-01B-06D-A270-10",
"TCGA-A6-6781-01B-06D-A274-01", "TCGA-A6-6781-01B-06D-A27A-05",

APPENDIX A. (cont.)

"TCGA-A6-6781-01B-06R-A277-07", "TCGA-A6-6781-01B-06R-A27D-13",
"TCGA-A7-A0DB-01C-02D-A272-09", "TCGA-A7-A0DB-01C-02R-A277-07",
"TCGA-A7-A0DB-01C-02R-A27D-13", "TCGA-A7-A13D-01B-04D-A272-09",
"TCGA-A7-A13D-01B-04R-A277-07", "TCGA-A7-A13D-01B-04R-A27D-13",
"TCGA-A7-A13E-01B-06D-A272-09", "TCGA-A7-A13E-01B-06R-A277-07",
"TCGA-A7-A13E-01B-06R-A27D-13", "TCGA-A7-A26E-01B-06D-A272-09",
"TCGA-A7-A26E-01B-06D-A275-01", "TCGA-A7-A26E-01B-06D-A27B-05",
"TCGA-A7-A26E-01B-06R-A277-07", "TCGA-A7-A26E-01B-06R-A27D-13",
"TCGA-A7-A26J-01B-02D-A272-09", "TCGA-A7-A26J-01B-02D-A275-01",
"TCGA-A7-A26J-01B-02D-A27B-05", "TCGA-A7-A26J-01B-02D-A27F-26",
"TCGA-A7-A26J-01B-02R-A277-07", "TCGA-A7-A26J-01B-02R-A27D-13",
"TCGA-B2-3923-01B-10D-A270-10", "TCGA-B2-3923-01B-10R-A277-07",
"TCGA-B2-3923-01B-10R-A27D-13", "TCGA-B2-3924-01B-03D-A270-10",
"TCGA-B2-3924-01B-03D-A274-01", "TCGA-B2-3924-01B-03D-A27A-05",
"TCGA-B2-3924-01B-03D-A27E-26", "TCGA-B2-3924-01B-03R-A277-07",
"TCGA-B2-3924-01B-03R-A27D-13", "TCGA-B2-5633-01B-04D-A270-10",
"TCGA-B2-5633-01B-04D-A274-01", "TCGA-B2-5633-01B-04D-A27A-05",
"TCGA-B2-5633-01B-04D-A27E-26", "TCGA-B2-5633-01B-04R-A277-07",
"TCGA-B2-5633-01B-04R-A27D-13", "TCGA-B2-5635-01B-04D-A270-10",
"TCGA-B2-5635-01B-04D-A274-01", "TCGA-B2-5635-01B-04D-A27A-05",
"TCGA-B2-5635-01B-04D-A27E-26", "TCGA-B2-5635-01B-04R-A277-07",
"TCGA-B2-5635-01B-04R-A27D-13", "TCGA-BK-A0CA-01B-02D-A272-09",
"TCGA-BK-A0CA-01B-02D-A275-01", "TCGA-BK-A0CA-01B-02D-A27B-05",
"TCGA-BK-A0CA-01B-02D-A27F-26", "TCGA-BK-A0CA-01B-02R-A277-07",
"TCGA-BK-A0CA-01B-02R-A27D-13", "TCGA-BK-A0CC-01B-04D-A272-09",
"TCGA-BK-A0CC-01B-04D-A275-01", "TCGA-BK-A0CC-01B-04D-A27B-05",
"TCGA-BK-A0CC-01B-04R-A277-07", "TCGA-BK-A0CC-01B-04R-A27D-13",
"TCGA-BK-A139-01C-08D-A272-09", "TCGA-BK-A139-01C-08D-A275-01",
"TCGA-BK-A139-01C-08D-A27B-05", "TCGA-BK-A139-01C-08D-A27F-26",
"TCGA-BK-A139-01C-08R-A277-07", "TCGA-BK-A139-01C-08R-A27D-13",
"TCGA-BK-A26L-01C-04D-A272-09", "TCGA-BK-A26L-01C-04D-A275-01",
"TCGA-BK-A26L-01C-04D-A27B-05", "TCGA-BK-A26L-01C-04D-A27F-26",
"TCGA-BK-A26L-01C-04R-A277-07", "TCGA-BK-A26L-01C-04R-A27D-13",
"TCGA-BL-A0C8-01B-04D-A271-08", "TCGA-BL-A0C8-01B-04D-A273-01",
"TCGA-BL-A0C8-01B-04D-A276-05", "TCGA-BL-A0C8-01B-04D-A27C-26",
"TCGA-BL-A0C8-01B-04R-A277-07", "TCGA-BL-A0C8-01B-04R-A27D-13",
"TCGA-BL-A13I-01B-04D-A271-08", "TCGA-BL-A13I-01B-04D-A276-05",
"TCGA-BL-A13I-01B-04R-A277-07", "TCGA-BL-A13I-01B-04R-A27D-13",
"TCGA-BL-A13J-01B-04D-A271-08", "TCGA-BL-A13J-01B-04D-A273-01",

APPENDIX A. (cont.)

```

"TCGA-BL-A13J-01B-04D-A276-05", "TCGA-BL-A13J-01B-04D-A27C-26",
"TCGA-BL-A13J-01B-04R-A277-07", "TCGA-BL-A13J-01B-04R-A27D-13")
tsb = setdiff(tsb, tsb[which(tsb %in% ffpe)])
}

# Find repeated samples
sampleID = substr(tsb, start = 1, stop = 15)
dp_samples = unique(sampleID[duplicated(sampleID)])

if(length(dp_samples)==0){
  message("ooo Not find any duplicated barcodes, return original input..")
  tsb
}else{
  uniq_tsb = tsb[! sampleID %in% dp_samples]
  dp_tsb = setdiff(tsb, uniq_tsb)

  add_tsb = c()

  # analyte = substr(dp_tsb, start = analyte_position, stop = analyte_position)
  # if analyte_target = "DNA"
  # analyte: D > G,W,X
  if(analyte_target == "DNA"){
    for(x in dp_samples){
      mulaliquots = dp_tsb[substr(dp_tsb,1,15) == x]
      analytes = substr(mulaliquots,
                        start = analyte_position,
                        stop = analyte_position)
      if(any(analytes == "D") & !(all(analytes == "D"))){
        aliquot = mulaliquots[which(analytes == "D")]
        add_tsb = c(add_tsb, aliquot)
        dp_tsb = setdiff(dp_tsb, mulaliquots)
      }
    }
  }else{
    for(x in dp_samples){
      mulaliquots = dp_tsb[substr(dp_tsb,1,15) == x]
      analytes = substr(mulaliquots,
                        start = analyte_position,
                        stop = analyte_position)

```

APPENDIX A. (cont.)

```
if(any(analytes == "H") & !(all(analytes == "H"))){
  aliquot = mulaliquots[which(analytes == "H")]
  add_tsb = c(add_tsb, aliquot)
  dp_tsb = setdiff(dp_tsb, mulaliquots)
} else if(any(analytes == "R") & !(all(analytes == "R"))){
  aliquot = mulaliquots[which(analytes == "R")]
  add_tsb = c(add_tsb, aliquot)
  dp_tsb = setdiff(dp_tsb, mulaliquots)
} else if(any(analytes == "T") & !(all(analytes == "T"))){
  aliquot = mulaliquots[which(analytes == "T")]
  add_tsb = c(add_tsb, aliquot)
  dp_tsb = setdiff(dp_tsb, mulaliquots)
}
}
}
# if analyte_target = "RNA"
# analyte: H > R > T
# else{
#
# }
if(length(dp_tsb) == 0){
  message("ooo Filter barcodes successfully!")
  c(uniq_tsb, add_tsb)
} else{
  # filter according to portion number
  sampleID_res = substr(dp_tsb, start=1, stop=15)
  dp_samples_res = unique(sampleID_res[duplicated(sampleID_res)])

  for(x in dp_samples_res){
    mulaliquots = dp_tsb[substr(dp_tsb,1,15) == x]
    portion_codes = substr(mulaliquots,
                          start = portion[1],
                          stop = portion[2])
    portion_keep = sort(portion_codes, decreasing = decreasing)[1]
    if(!all(portion_codes == portion_keep)){
      if(length(which(portion_codes == portion_keep)) == 1){
        add_tsb = c(add_tsb, mulaliquots[which(portion_codes == portion_keep)])
        dp_tsb = setdiff(dp_tsb, mulaliquots)
      } else{
```

APPENDIX A. (cont.)

```
dp_tsb = setdiff(dp_tsb, mulaliquots[which(portion_codes != portion_keep)])
}
}
}

if(length(dp_tsb)==0){
  message("ooo Filter barcodes successfully!")
  c(uniq_tsb, add_tsb)
} else{
  # filter according to plate number
  sampleID_res = substr(dp_tsb, start=1, stop=15)
  dp_samples_res = unique(sampleID_res[duplicated(sampleID_res)])
  for(x in dp_samples_res){
    mulaliquots = dp_tsb[substr(dp_tsb,1,15) == x]
    plate_codes = substr(mulaliquots,
                        start = plate[1],
                        stop = plate[2])
    plate_keep = sort(plate_codes, decreasing = decreasing)[1]
    add_tsb = c(add_tsb, mulaliquots[which(plate_codes == plate_keep)])
    dp_tsb = setdiff(dp_tsb, mulaliquots)
  }

  if(length(dp_tsb)==0){
    message("ooo Filter barcodes successfully!")
    c(uniq_tsb, add_tsb)
  } else{
    message("ooo barcodes ", dp_tsb, " failed in filter process, other barcodes will be returned.")
    c(uniq_tsb, add_tsb)
  }
}
}

#-----
# Test and Check result
#-----
```

APPENDIX A. (cont.)

SNP cases replicate filter

```
tsb_maf = cases_maf_paired
```

```
# Sort replicate filter
```

```
require(tidyverse)
```

```
test_tsb_maf <- tsb_maf %>% substr(1,15) %>% unique()
```

```
uniq_tsb_maf = tcga_replicateFilter(tsb = tsb_maf, analyte_target = "DNA", filter_FFPE = TRUE, full_barcode = TRUE)
```

```
all(test_tsb_maf %in% substr(uniq_tsb_maf, 1, 15))
```

```
all.equal(sort(test_tsb_maf), sort(substr(uniq_tsb_maf, 1, 15)))
```

Expression cases replicate filter

```
tsb_exp = cases_exp_paired
```

```
# Sort replicate filter
```

```
require(tidyverse)
```

```
test_tsb_exp <- tsb_exp %>% substr(1,15) %>% unique()
```

```
uniq_tsb_exp = tcga_replicateFilter(tsb = tsb_exp, analyte_target = "RNA", filter_FFPE = TRUE, full_barcode = TRUE)
```

```
all(test_tsb_exp %in% substr(uniq_tsb_exp, 1, 15))
```

```
all.equal(sort(test_tsb_exp), sort(substr(uniq_tsb_exp, 1, 15)))
```

CNV cases replicate filter

```
tsb_cnv = cases_cnv_paired
```

```
# Sort replicate filter
```

```
require(tidyverse)
```

```
test_tsb_cnv <- tsb_cnv %>% substr(1,15) %>% unique()
```

```
uniq_tsb_cnv = tcga_replicateFilter(tsb = tsb_cnv, analyte_target = "DNA", filter_FFPE = TRUE, full_barcode = TRUE)
```

APPENDIX A. (cont.)

```
all(test_tsb_cnv %in% substr(uniq_tsb_cnv, 1, 15))
all.equal(sort(test_tsb_cnv), sort(substr(uniq_tsb_cnv, 1, 15)))
```

R Codes of Summarizing of Clinical Data

```
library(TCGAbiolinks)
library(SummarizedExperiment)
library(dplyr)
library(DT)

cancer <- "LUAD"

common_patients_paired <- read.csv(file = paste0(cancer, "_common_patients_paired.tsv"), header =
FALSE, sep = "\t", stringsAsFactors=FALSE)
common_patients_paired <- as.character(unlist(common_patients_paired))

common_patients <- read.csv(file = paste0(cancer, "_common_patients.tsv"), header = FALSE, sep =
"\t")
common_patients <- as.character(unlist(common_patients))

clinic <- GDCquery_clinic(project = paste0("TCGA-", cancer), type = "clinical")

clinic_total <- clinic[clinic[, "bcr_patient_barcode"] %in% substr(common_patients, 1, 12), ]
clinic_paired <- clinic[clinic[, "bcr_patient_barcode"] %in% substr(common_patients_paired, 1, 12), ]

# Clinical data of patients who have paired data
clinic_paired %>% select(gender,age_at_diagnosis,tumor_stage,vital_status) %>% head
table(clinic_paired$gender)
table(clinic_paired$tumor_stage)
table(clinic_paired$vital_status)
summary(clinic_paired$age_at_diagnosis)

# Clinical data of all patients who have all types of data
clinic_total %>% select(gender,age_at_diagnosis,tumor_stage,vital_status) %>% head
table(clinic_total$gender)
table(clinic_total$tumor_stage)
table(clinic_total$vital_status)
summary(clinic_total$age_at_diagnosis)
```

CIRCULUM VITAE

Personal Information

Talip ZENGİN

TURKISH (T.C.)

04 February 1987, Milas

E-mail: talipzengin@mu.edu.tr

Education

Degree	Institution	Graduation year
High School	Milas Anatolian High School	2005
Bachelor	Izmir Institute of Technology Department of Molecular Biology and Genetics	2005-2011
Master of Science	Izmir Institute of Technology Department of Molecular Biology and Genetics	2011-2014
Master of Science	Muğla Sıtkı Koçman University Department of Bioinformatics	2016-2019

Publication

Peer Review Publication

Zengin Talip, Ekinici Burcu, Küçükköse Cansu, Yalçın Özuysal Özden. 2015. IRF6 is Involved in the Regulation of Cell Proliferation and Transformation in MCF10A Cells Downstream of Notch Signaling. Plos ONE.

International Symposiums

Zengin Talip, Süzek Tuğba, 2018. TCGA Lung Cancer Analysis Pipeline. Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics – BCB, USA.

Zengin T., Yalçın Özuysal Ö. 2013. IRF6 is a downstream mediator of Notch signaling in breast epithelial cells. VII. Notch Meeting, Athens, Greece.

Zengin T., Yalçın Özuysal Ö. 2013. IRF6 is a downstream mediator of Notch signaling in breast epithelial cells. II. International Congress of the Molecular Biology Association of Turkey, Istanbul, Turkey.

Book Chapter

Balta E., Boz S., Rende U., Zengin T., Erbaşol I. 2012. Biyoenformatik I - Dizi Kıyaslamaları. Nobel Akademik Yayıncılık. Ankara.

