T. ZENGİN DEPARTMENT OF BIOINFORMATICS DOCTORAL THESIS MUĞLA 2023

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

DEPARTMENT OF BIOINFORMATICS

DETECTION OF IMMUNE TARGETABLE CANCER BIOMARKERS IN THE CANCER GENOME ATLAS DATASETS

DOCTORAL (Ph.D.) THESIS

TALİP ZENGİN

JUNE 2023

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

DEPARTMENT OF BIOINFORMATICS

DETECTION OF IMMUNE TARGETABLE CANCER BIOMARKERS IN THE CANCER GENOME ATLAS DATASETS

DOCTORAL (Ph.D.) THESIS

TALİP ZENGİN

JUNE 2023

MUGLA SITKI KOÇMAN UNIVERSITY

Graduate School of Natural and Applied Sciences

APPROVAL OF THE THESIS

The thesis submitted by **TALİP ZENGİN** with the title of "**DETECTION OF IMMUNE TARGETABLE CANCER BIOMARKERS IN THE CANCER GENOME ATLAS DATASETS**" has been unanimously accepted by the jury members on the 5th of June 2023 to fulfill the requirements for the degree of Doctor of Philosophy in the Department of Bioinformatics.

THESIS JURY MEMBERS

Assoc. Prof. Dr. Barış SÜZEK **(Head of Jury)** Signature: ______________
Department of Computer Engineering,
Muğla Sıtkı Koçman University

Assist. Prof. Dr. Tuğba SÜZEK **(Supervisor)** Signature: ______________
Department of Computer Engineering,
Muğla Sıtkı Koçman University

Prof. Dr. Özden YALÇIN ÖZUYSAL **(Member)** Signature: ______________
Department of Molecular Biology and Genetics,
Izmir Institute of Technology

Assoc. Prof. Dr. Can KÜÇÜK **(Member)** Signature: ______________
Department of Medical Biology,
Dokuz Eylül University

Assist. Prof. Dr. Yavuz OKTAY **(Member)** Signature: ______________
Department of Medical Biology,
Dokuz Eylül University

APPROVAL OF HEAD OF THE DEPARTMENT

Assist. Prof. Dr. Tuğba SÜZEK, **Head of Department** Signature: ______________
Department of Computer Engineering,
Muğla Sıtkı Koçman University

Assist. Prof. Dr. Tuğba SÜZEK, **Supervisor** Signature: ______________
Department of Computer Engineering,
Muğla Sıtkı Koçman University

Defense Date: 05/06/2023

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
05/06/2023

ABSTRACT

DETECTION OF IMMUNE TARGETABLE CANCER BIOMARKERS IN THE CANCER GENOME ATLAS DATASETS

Talip ZENGİN

Doctoral (Ph.D.) Thesis
Graduate School of Natural and Applied Sciences
Department of Bioinformatics
Supervisor: Assist. Prof. Dr. Tuğba SÜZEK
JUNE 2023, 144 pages

Advances in bioinformatics data analysis have enabled the identification of genomic, transcriptomic, and epigenetic variations in tumor samples, that can be used for personalized cancer treatment. Cancer immunotherapy has been developed specifically as an advanced cancer therapy, involved in the interactions between cancer cells and immune cells in the tumor microenvironment, which are the basis of cancer cell for immune escape. The efficacy of immunotherapy can be affected by genomic alterations, tumor neoantigens, immune phenotype, and other biomarkers in the tumor microenvironment. Technological advancements have permitted more in-depth investigation into the link between cancer and the immune system, which has led to advancements in biomarker discovery. Since The Cancer Genome Atlas (TCGA) project provides both molecular data including simple nucleotide variations (SNVs), copy number variations (CNVs), DNA methylation, gene expression, miRNA expression, and clinical data such as drug responses and survival data, these data can be used for many studies such as biomarker detection and drug response. In this project, we focused on tumor-associated proteins through integrative clusters to analyze predictive and prognostic markers to compare clinical variables and characteristics of patient clusters. Finally, we developed a web tool that assists users in querying data and visualizing results specific to cancer, cohorts, and biomarkers. The web tool provides an interface to combine candidate biomarkers which affect survival statistics of patient clusters/sub-cohorts.

Keywords: Cancer, immunotherapy, integrative clustering, iCluster, TCGA

ÖZET

KANSER GENOM ATLASI VERİ SETLERİNDE BAĞIŞIKLIKLA HEDEFLENEBİLEN KANSER BİYOBELİRTEÇLERİNİN TESPİTİ

Talip ZENGİN

Doktora Tezi
Fen Bilimleri Enstitüsü
Biyoinformatik Anabilim Dalı
Danışman: Dr. Öğretim Üyesi Tuğba SÜZEK
HAZİRAN 2023, 144 sayfa

Biyoinformatik veri analizindeki gelişmeler, tümör örneklerinde kişiselleştirilmiş kanser tedavisi için kullanılabilecek genomik, transkriptomik ve epigenetik varyasyonların tanımlanmasını sağlamıştır. Kanser immünoterapisi, özellikle kanser hücreleri ile tümör mikroçevresindeki immün hücreler arasındaki etkileşimlerde yer alan ve bağışıklık sisteminden kaçışı durduran ileri bir kanser tedavisi olarak geliştirilmiştir. İmmünoterapinin etkinliği, genomik değişiklikler, tümör neo-antijenleri, immün fenotip ve tümör mikroçevresindeki diğer biyobelirteçlerden etkilenebilir. Teknolojik gelişmeler, kanser ve bağışıklık sistemi arasındaki bağlantının daha derinlemesine araştırılmasına izin vermiş ve bu da biyobelirteç keşfinde ilerlemelere yol açmıştır. Kanser Genom Atlası (TCGA) projesi hem basit nükleotid varyasyonları (SNV'ler), kopya sayısı varyasyonları (CNV'ler), DNA metilasyonu, gen ifadesi, miRNA ifadesi gibi moleküler verileri hem de ilaç yanıtları ve sağ kalım verileri gibi klinik verileri sağladığından, bu veriler biyobelirteç tespiti ve ilaç yanıtı gibi birçok çalışma için kullanılabilir. Bu projede, hasta kümelerinin klinik değişkenlerini ve özelliklerini karşılaştırmak için öngörücü ve prognostik belirteçleri analiz etmek üzere bütünleştirici kümeler aracılığıyla tümörle ilişkili proteinlere odaklandık. Son olarak, kansere, hasta gruplarına ve biyobelirteçlere özgü sonuçları sunan ve görselleştiren bir web aracı geliştirdik. Web aracı, hasta kümelerinin/alt hasta gruplarının sağ kalım istatistiklerini etkileyen aday biyobelirteçleri birleştirmek için bir ara yüz olarak kullanılabilir.

Anahtar Kelimeler: Kanser, immunoterapi, bütünleştirici kümeleme, TCGA

In memory of the cancer patients, we lost

ACKNOWLEDGEMENTS

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

I am very grateful to Prof. Dr. Özden YALÇIN ÖZUYSAL and Assoc. Prof. Dr. Barış Ethem SÜZEK for their contributions and comments on my thesis.

I would also like to thank Assoc. Prof. Dr. Can KÜÇÜK and Assist. Prof. Dr. Yavuz OKTAY for evaluating my thesis.

I would like to thank all my friends and colleagues who supported my Ph.D. thesis and my training in bioinformatics.

I would like to thank Mehtap ÖZTÜRK for her endless support throughout my thesis. I would like to thank Bagemon for keeping me company while I was writing my thesis. I would also like to thank Ezgi BURGAN for taking care of Bagemon in my absence. I would especially like to thank Özgür KARATAŞ for encouraging me since the beginning of my long academic career.

I would like to express my deepest gratitude to my family for their love, support and prayers throughout my life.

I would like to thank Turkish National Institutes of Health (TÜSEB) and The Scientific and Technological Research Council of Türkiye (TÜBİTAK) because the computational analyses reported in this thesis were performed on two different servers sponsored by TÜSEB (4583) and TÜBİTAK (20AG006) projects.

TABLE OF CONTENTS

ABSTRACT .... **iv**

ÖZET .... **v**

ACKNOWLEDGEMENTS .... **vii**

TABLE OF CONTENTS .... **viii**

LIST OF TABLES .... **x**

LIST OF FIGURES .... **xi**

1. CHAPTER: INTRODUCTION **1**

2. CHAPTER: Comprehensive Profiling of Genomic and Transcriptomic Differences Between Risk Groups of Lung Adenocarcinoma and Lung Squamous Cell Carcinoma **21**

2.1. Abstract 21

2.2. Methods for Risk Clustering 22

2.3. Results of Risk Clustering 24

2.4. Discussion of the method 27

3. CHAPTER: Integrative Clustering and Comparison of Clusters for Immune Targetable Cancer Biomarkers **29**

3.1. Introduction 29

3.1.1. Integrative clustering 30

3.2. Methods 34

3.2.1. Data 34

3.2.2. Preparation of data for integrative clustering 35

3.2.3. Integrative clustering analysis 36

3.2.4. Feature selection with gene and pathway enrichment 37

3.2.5. Curated TCGA data from publications 37

3.2.6. Statistical Analysis 38

3.2.7. Analysis of the known targets and the approved/used immunotherapeutics 38

3.2.8. Integrating the iCluster data into the web tool 38

3.3. Results 39

3.3.1. Integrative clustering of LUAD patients 39

3.3.1.1. Feature selection 43

3.3.1.2. Pathway Enrichment 43

3.3.1.3. Comparison of LUAD iClusters for clinical variables and sample biology 44

3.3.2. Integrative clustering of LUSC patients 56

3.3.2.1. Feature selection 60

3.3.2.2. Pathway Enrichment 60

3.3.2.3. Comparison of LUSC iClusters for clinical variables and sample biology 61

3.4. Discussion 71

4. CHAPTER: TCGAnalyzeR: A Web Application for Integrative Visualization of Molecular and Clinical Data of Cancer Patients for Cohort and Associated Gene Discovery 75

4.1. Abstract 75

4.2. Methods 76

4.2.1. Data 76

4.2.2. SNV Analysis 76

4.2.3. CNV Analysis 76

4.2.4. Differential Expression Analysis (DEA) 77

4.2.5. Pre-computed Patient Clusters and Sample Subtypes 77

4.2.6. Survival Analysis 78

4.2.7. Visualization 78

4.3. Examples of Data Visualization from TCGAnalyzeR 79

4.4. Conclusion 84

REFERENCES 85

APPENDIX A. 94

APPENDIX B. 123

CURRICULUM VITAE 129

LIST OF TABLES

Table 1.1. Approved antibody therapeutics used for cancer in clinic. 13
Table 1.2. Antibody therapeutics given to patients in TCGA projects. 18
Table 1.3. Targetable and targeted antigens in TCGA projects. 19
Table 3.1. Number of patients with tumor samples by cancer types for five different data types. 34

LIST OF FIGURES

Figure 1.1. Mechanisms of tumor cell killing by antibodies.. 4

Figure 1.2. Therapeutic antibody types and target proteins 6

Figure 1.3. Mechanisms of immune reaction to target cell and possible antibody therapeutics.. 8

Figure 2.1. Flowchart of the methods used for risk clustering and analysis of genomic and transcriptomic variations between risk groups 24

Figure 2.2. Gene expression signature and risk clustering of LUAD test dataset 25

Figure 2.3. Gene expression signature and risk clustering of LUSC test dataset 26

Figure 3.1. Integration of diverse data types by a latent variable approach............... 31

Figure 3.2. The proposed Bayesian integrative clustering framework 32

Figure 3.3. The flowchart summarizing the methods used for integrative clustering and comparison of clusters for immune targetable cancer biomarkers... 39

Figure 3.4. Bayesian Information Criteria (BIC) and the deviance ratio vs k for LUAD dataset.. 40

Figure 3.5. Posterior probabilities of genomic features that drive the integrative clustering.. 41

Figure 3.6. Heatmap showing LUAD samples (columns) and genomic features (rows) 42

Figure 3.7. Pathways of SNVs, gene expression and miRNA expression used for integrative clustering of LUAD tumor samples.................................... 44

Figure 3.8. Relationships between iClusters and clinical variables of LUAD patients 45

Figure 3.9. Survival and interval analysis results of LUAD iClusters....................... 46

Figure 3.10. Heatmap comparison of LUAD iClusters with TCGA subtypes curated from publications in literature.. 47

Figure 3.11. Relationships between LUAD iClusters and TCGA subtypes 48

Figure 3.12. Relationships between LUAD iClusters and immune subtypes 49

Figure 3.13. Comparison of LUAD iClusters for tumor infiltrated immune cells..... 50

Figure 3.14. Comparison of LUAD iClusters for Tumor Mutation Burden and Homologous DNA Repair Deficiency 51

Figure 3.15. Correlation matrix between TMB/HRD and tumor infiltrated immune cells in LUAD tumor samples 51

Figure 3.16. Expression heatmap of tumor associated proteins targetable by antibody therapeutics in LUAD tumor samples 52

Figure 3.17. Expression heatmap of immune associated proteins targetable by antibody therapeutics in LUAD tumor samples 53
Figure 3.18. Comparison of LUAD iClusters for expression of targetable proteins by antibody therapeutics 54
Figure 3.19. Relationships between iClusters and risk clusters of LUAD patients ... 55
Figure 3.20. Comparison of LUAD risk clusters for expression of targetable proteins by antibody therapeutics 56
Figure 3.21. Bayesian Information Criteria (BIC) and the deviance ratio vs k for LUSC dataset 57
Figure 3.22. Posterior probabilities of genomic features that drive the integrative clustering.. 58
Figure 3.23. Heatmap showing LUSC samples (columns) and genomic features (rows) 59
Figure 3.24. Pathways of SNVs, gene expression and miRNA expression used for integrative clustering of LUSC tumor samples.................................... 61
Figure 3.25. Relationships between iClusters and clinical variables of LUSC patients 62
Figure 3.26. Survival and interval analysis results of LUSC iClusters..................... 63
Figure 3.27. Heatmap comparison of LUSC iClusters with TCGA subtypes in literature 63
Figure 3.28. Relationships between LUSC iClusters and TCGA subtypes 64
Figure 3.29. Comparison of LUSC iClusters for tumor infiltrated immune cells...... 65
Figure 3.30. Comparison of LUSC iClusters for Tumor Mutation Burden and Homologous DNA Repair Deficiency 66
Figure 3.31. Correlation matrix between TMB/HRD and tumor infiltrated immune cells in LUSC tumor samples 66
Figure 3.32. Expression heatmap of tumor associated proteins targetable by antibody therapeutics in LUSC tumor samples 67
Figure 3.33. Expression heatmap of immune associated proteins targetable by antibody therapeutics in LUSC tumor samples 68
Figure 3.34. Comparison of LUSC iClusters for expression of targetable proteins by antibody therapeutics 69
Figure 3.35. Relationships between iClusters and risk clusters of LUSC patients.... 70
Figure 3.36. Comparison of LUSC risk clusters for expression of targetable proteins by antibody therapeutics 71
Figure 4.1. Flowchart of methods used for preparing the data for TCGAnalyzeR web tool 78

Figure 4.2. Bar graph summarizing numbers of samples, genes with SNVs, genes with CNVs and Differentially Expressed Genes (DEGs) at main page of TCGAnalyzeR web tool.. 79

Figure 4.3. Oncoplot of genes with SNVs predicted as driver by SomInaClust........ 79

Figure 4.4. Bar graph and CNV analysis results of genomic regions with CNVs in LUAD dataset 80

Figure 4.5. Volcano plot and analysis results of differentially expressed genes of LUAD samples 81

Figure 4.6. Pathway enrichment of differentially expressed genes in LUAD tumor samples.. 82

Figure 4.7. EPCAM gene expression in tumor and adjacent normal samples of LUAD 82

Figure 4.8. Pie charts and survival analysis of pathologic stage and existence of synchronous malignancy.. 83

1. CHAPTER

INTRODUCTION

Somatic DNA mutations and epigenetic alterations that change transcriptional levels, protein function, and cell phenotype are the hallmarks of cancer. Therefore, molecular data such as mutations, gene expression, or DNA methylation can provide insight into the mechanism underlying carcinogenesis and progression. At the level of genomic and epigenomic alterations, multi-omics cancer analysis provides a broader understanding of these disorders (Sengupta et al., 2018; Xu et al., 2019). During carcinogenesis and progression, genomic variations produced by DNA copy number variations (CNVs), or simple nucleotide variations (SNVs) are prominent (Li et al., 2017; Ren et al., 2019; Yang et al., 2017). The variability of cancer phenotype is also influenced by epigenetic regulations of DNA methylation in the cancer genome (Yang et al., 2020). Omics studies reveal the enormous variability of genomic and epigenomic dysregulation, notably in some malignancies, and copy number variation (CNV) plays a significant regulatory role in both transcriptional abnormalities and tumorigenesis. (Hull et al., 2020).

For the integrated analysis of numerous forms of -omics data, many methodologies and findings have been published (Kristensen et al., 2014; Richardson et al., 2016). For instance, Shen et al. created the integrative clustering approach (iCluster) to identify the molecular subgroups of colon, lung, and breast malignancies (Shen et al., 2009). DNA methylation, an epigenetic marker extensively investigated in the cancer genome, is widely associated with various types of cancer. Studies have revealed that cancer often exhibits both reduced and increased DNA methylation on a genome-wide scale, particularly in specific gene promoters (Irizarry et al., 2009). The regulation of gene expression is frequently intertwined with epigenetic factors such as chromatin structure and DNA methylation. To comprehensively comprehend the significance of DNA methylation in cancer, it is essential to investigate the connection between DNA methylation and gene expression (Fleischer et al., 2014; Gevaert et al., 2015). Research has demonstrated an inverse relationship between DNA methylation in promoter regions and gene expression, indicating that higher levels of DNA methylation in these regions are associated with lower gene expression. Conversely, DNA methylation in gene bodies has been found to have a positive correlation with gene expression, suggesting that increased DNA methylation within these regions is linked to higher levels of gene expression (Shen et al., 2013).

Research on cancer types is used to identify cancer biomarkers, to understand the tumor formation process, or to characterize cancer subgroups that show different clinical results using integrated biological and clinical data of a group of patients with a particular type of cancer. Genome analysis of a cancer patient is performed to identify genomic variations that can be used for specially designed cancer treatment (Vazquez et al., 2012). Information obtained from cancer genome analysis on cancer type can be used in determining the patient's cancer type and analyzing the biomarkers of the tumor. Therefore, the best chemical components can be determined specifically for a patient. In ordinary clinical practice, patients are grouped according to their general characteristics (age, smoking status, and performance status) and tumor (tumor size, histological stage, expression of several proteins, and metastasis status). However, such grouping does not take into account the molecular characteristics of the tumor in depth. Therefore, methods that have recently become available such as gene expression profiling and various genomic tests have been developed to determine treatment and predict clinical outcomes (Cardoso et al., 2016). Personalized cancer therapy needs a strong knowledge of cancer genomes, expertise with cancer research analytical methodologies, knowledge of targeted therapeutic action mechanisms, and strategies for organizing and comprehending large data sets.

Identifying medications projected to be effective based on the genetic profile of a patient's tumor is perhaps the most difficult task (Schilsky et al., 2014).

Despite advances in cancer treatment, there are subgroups of patients who do not respond to traditional chemotherapy treatments or whose disease has relapsed. Recently, researchers focused on the role of the immune system in cancer control. The escape of cancer cells from the immune response is an important sign of cancer (Hanahan and Weinberg, 2011; Beck and Blanpain, 2013). It is also known that the interaction between immune cells and cancer cells in microenvironment of the tumor is the basis for the cancer cell escaping immune response. To solve this problem, cancer immunotherapy has been developed as a treatment method for various malignancies. Cancer immunotherapy involves the development of strategies that leverage the mechanisms underlying the interaction between immune cells and cancer cells within the tumor microenvironment. Until now, more than 100 monoclonal antibodies (mAbs) have been approved by the US Food and Drug Administration (FDA) for treatment of cancer and other various human diseases. These antibodies possess the ability to selectively bind to specific antigens and elicit cytotoxic effects by either neutralizing them or triggering programmed cell death (Jin et al., 2022). Additionally, they have the capability to stimulate innate immune responses, complement-dependent cytotoxicity (CDC), antibody-dependent cellular cytotoxicity (ADCC), and antibody-dependent cellular phagocytosis (ADCP) (Figure 1.1).

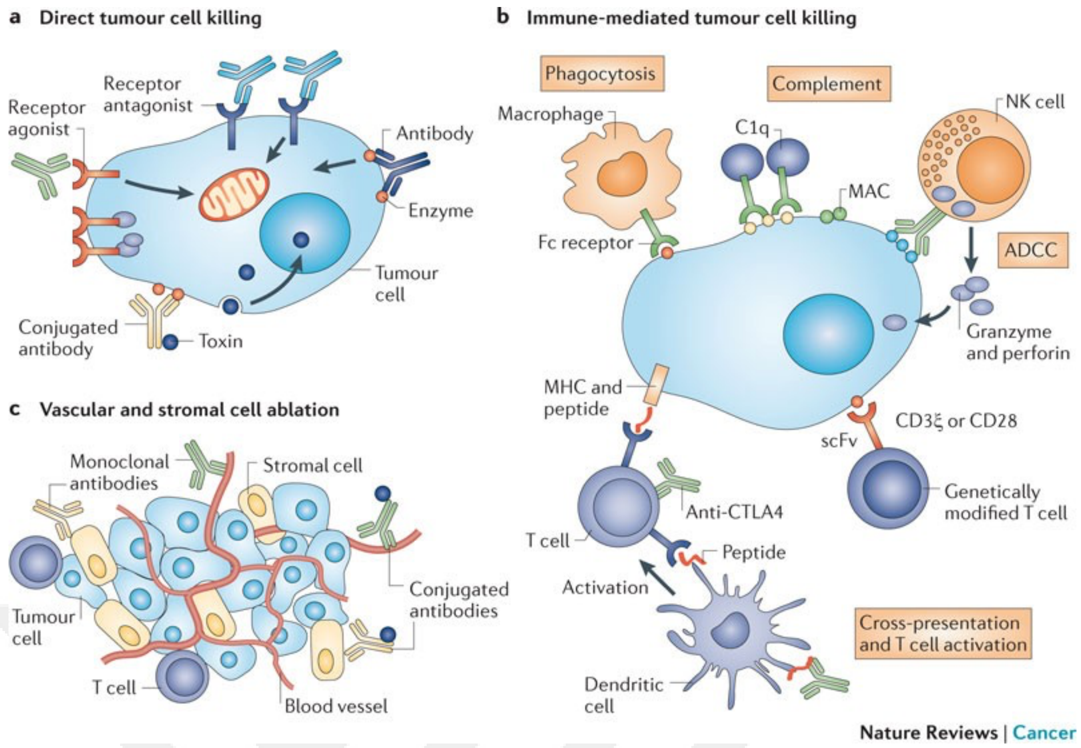


Figure 1.1. Mechanisms of tumor cell killing by antibodies. (Taken from Scott et al., 2012)

Expanding on the accomplishments of immunoglobulin G (IgG) mAbs, alternative therapeutic options have gained recognition and started to be utilized. These options include other antibody formats such as antibody fragments, non-IgG scaffold proteins, and bispecific antibodies (BsAbs), and as well as antibody derivatives like antibody-drug conjugates (ADCs) and immunocytokines (Figure 1.2) for a wide range of cancers (Jin et al., 2022). The effectiveness of an antibody in combating tumors can be significantly enhanced by attaching a potent cytotoxic small molecule to the antibody, resulting in antibody-drug conjugates (ADCs). ADCs have the capacity to selectively transport potent small-molecule drugs directly to cancer cells, ensuring targeted and precise delivery. of interest, leading to their programmed cell death (Figure 1.2). The FDA has already approved more than 10 ADCs for the treatment of cancer, and over 80 ADCs are currently being investigated in clinical trials. In addition, antibodies can also be conjugated with various other types of molecules such as radionuclides, protein toxins and oligonucleotides (Jin et al., 2022).

Bispecific antibodies (BsAbs) possess the ability to bind simultaneously to two distinct antigens. A bispecific antibody (BsAb) known as a bispecific T cell engager is characterized by having one arm that specifically targets CD3 on T cells and another arm that recognizes specific proteins present on tumor cells. This interaction activates T cells, enabling them to eliminate the malignant cells (Jin et al., 2022). An example of a pioneering BsAb called blinatumomab, which can bind to CD19 and CD3, received FDA approval in 2014 for treating B cell precursor acute lymphoblastic leukemia. In addition to their binding ability to T cells, bispecific antibodies (BsAbs) have been engineered to bind different immune cells, including natural killer (NK) cells and macrophages, in the context of cancer therapy (Figure 1.2). Another innovative category of antibody-based immunotherapies is antibody-cytokine fusion proteins, known as immunocytokines. A variety of very tiny proteins known as cytokines are essential for controlling immune responses. Proinflammatory cytokines can be administered systemically, however doing so frequently causes significant off-target damage, which limits their effective dose and therapeutic potential. Immunocytokines have been developed to address this issue. Cytokines combine with antibodies or antibody fragments to form these immunocytokines. Through this fusion, off-target toxicity is reduced and the therapeutic effectiveness of immunomodulatory cytokines such as IL-2, IL-12, and TNF in the tumor microenvironment is increased (Figure 1.2). This approach aims to activate anti-cancer immune responses while minimizing systemic side effects (Jin et al., 2022).

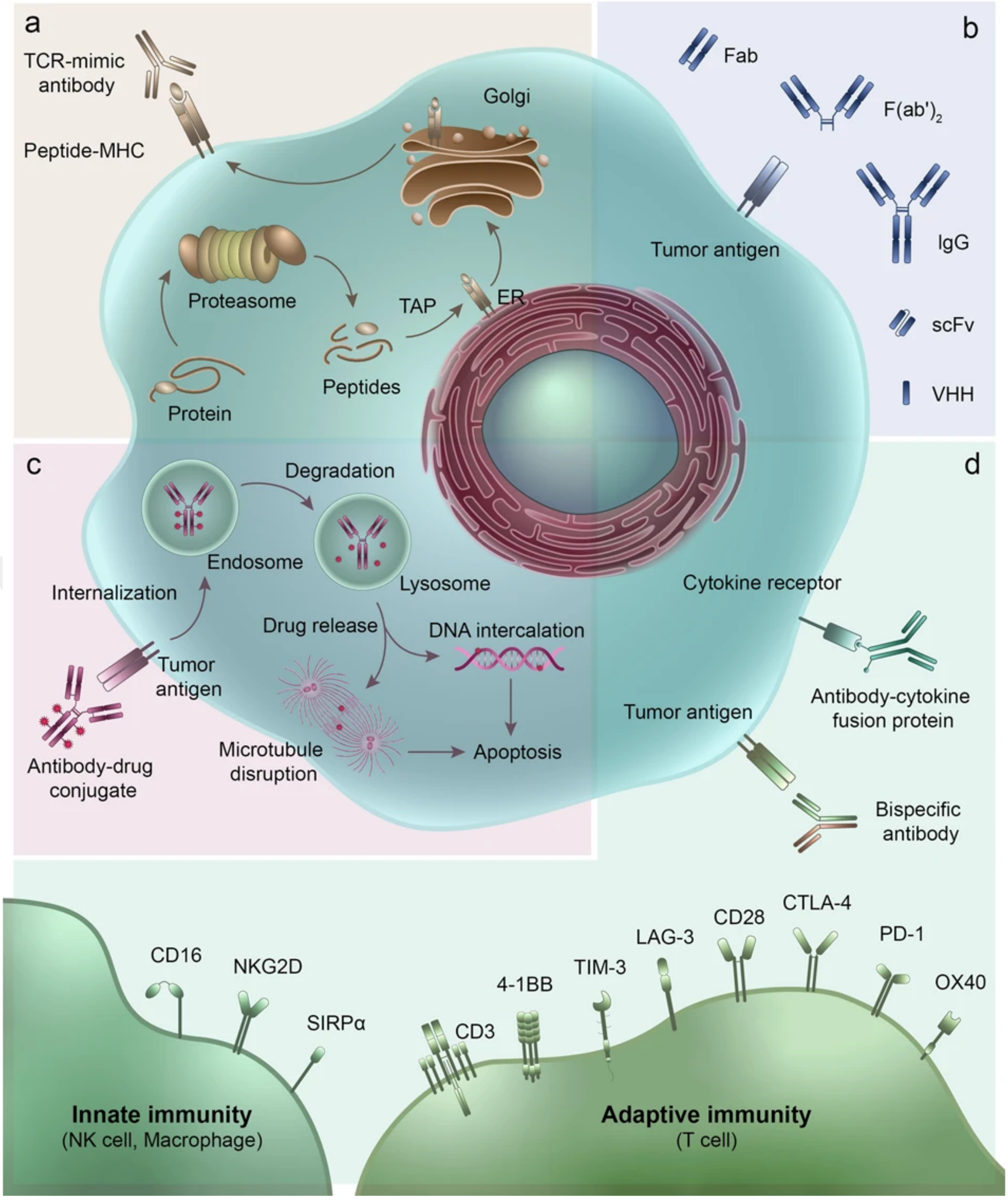


Figure 1.2. Therapeutic antibody types and target proteins. a) TCR-mimic antibody; **b)** IgG and antibody fragments; **c)** Antibody-drug conjugate (ADC); **d)** Bispecific antibody and Antibody-cytokine fusion protein. (Taken from Jin et al., 2022)

Monoclonal antibodies (mAbs) typically bind to antigens on the surface of cells, whereas many cancer-associated proteins are located inside the cell. On the other hand, T cell receptors (TCRs) can recognize small parts of intracellular proteins presented by major histocompatibility complex (MHC). To target proteins of interest within tumor cells or other cells, researchers are using antibodies that mimic the epitope-recognition segment of TCRs, known as TCR mimic (TCRm) antibodies (Figure 1.2). These TCRm antibodies combine the ability to target pMHC with the robustness typically found in IgG mAbs, offering improved potential for therapeutic applications (Jin et al., 2022).

In the past ten years, there has been a significant shift towards antibody therapeutics which can induce immune response against cancer cells. Immune-checkpoint inhibitors (ICI) which are first generation of antibody immunotherapies, block interactions between receptors and ligands involved in dampening the activation or function of T cells. Although ICIs have notable benefits for treatment of many cancer types, some patients cannot benefit from ICIs. Certain types of tumor immune microenvironments (TIME) have higher chance to respond to ICIs. By going deeper into the complexity of the TIME, advanced biomarkers can be discovered to identify patient clusters which show better response to ICIs (Binnewies et al., 2018).

It is crucial to predict the responsiveness to ICIs based on analysis of tumor infiltrated immune cell composition and their activity states with their receptor repertoires. Existing research has primarily focused on establishing tumor microenrivonment data using available techniques such as immunohistochemistry or RNAseq of bulk tissue. CIBERSORT (Newman et al., 2015) and XCell (Aran et al., 2017) can be used to estimate the fractions of tumor infiltrated immune cells using bulk RNAseq data. Immunoscore (Bindea et al., 2013), on the other hand, combines immuno-histochemistry and gene expression to predict outcome of clustered patients. Exploring new techniques such as single cell sequencing and advancing the resolution of TIME data will be essential to predict effectiveness of current therapies and to develop future immunotherapies (Binnewies et al., 2018).

Immune checkpoint proteins have the main function to prevent autoimmunity and tissue damage during pathogenic infections. These proteins are inhibitory receptors expressed on the surface of T cells and tissue cells and prevent T cells from attacking tissue cells (Pardoll et al., 2012). In many malignancies, immune checkpoint inhibitors (ICIs) treatment brought a new opportunity, with considerable survival improvements. Antibodies against Cytotoxic T-lymphocyte protein 4 (CTLA4), Programmed Cell Death 1 (PD-1) or Programmed Cell Death-Ligand 1 (PD-L1) have been used as second or first treatment option in solid tumors (Figure 1.3). Despite advancements in clinical therapy with ICIs, most of the patients cannot give response them. Only 40–45% of the melanoma patients show response to first line nivolumab or pembrolizumab in first-line treatment and 20% of non-small cell lung cancer (NSCLC) patients in second-line treatment (Robert et al., 2015; Garon et al., 2015; Brahmer et al., 2015).

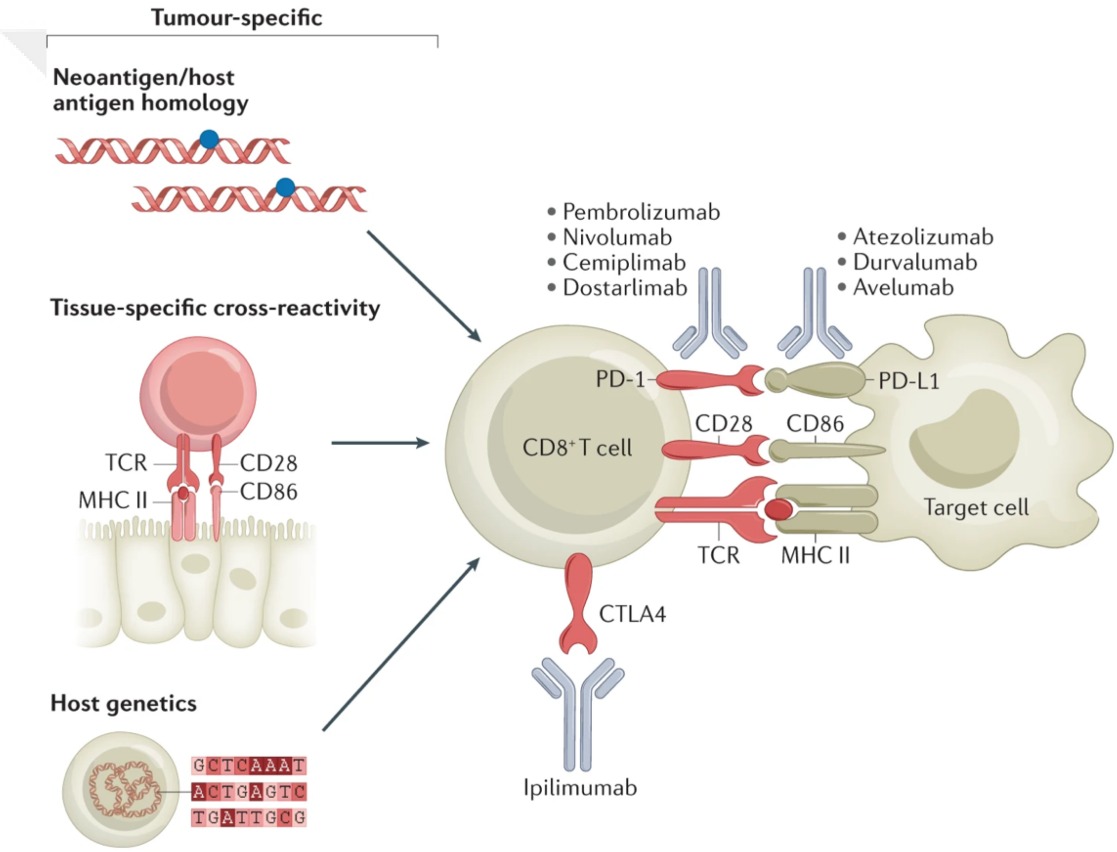


Figure 1.3. Mechanisms of immune reaction to target cell and possible antibody therapeutics. Interaction of T cells with malignant cells, and the mechanisms of immune-checkpoint blockade. (Adapted from Johnson et al., 2022)

More research has been conducted in recent years to develop predictive biomarkers for the effectiveness of ICIs, and a thorough understanding of tumor biomarkers has been acquired. Understanding of the genomic changes, tumor neoantigens, immune phenotype in the tumor microenvironment, and liquid biopsy biomarkers have made many new advances in this area. Various biomarker discovery methodologies have evolved as such as multiplex immunohistochemistry and single cell sequencing. The predictive biomarkers aid in the discovery of ICI therapeutic mechanisms and processes of tumor-host immune interaction in order to better understand tumor prognosis and drug response/resistance mechanisms (Bai et al., 2020).

PD-1/PD-L1 Expression: Tumor cells that produce the PD-L1 protein can block CD8+ "killer" T lymphocytes from destroying tumors by turning them off (through the PD-1 receptor). The presence of PD-L1 in tumors is utilized as the major biomarker for determining whether individuals may benefit from these immunotherapies. Patients who have tumor with strong expression of PD-L1 can react (Taube et al., 2014).

Tumor Infiltrated Lymphocytes: The fraction of tumor infiltrated CD8+ T cells within tumors estimated by Immunoscore (Angell and Galon, 2013), is linked to improved outcomes in cancer patients, independently of therapy. To protect themselves from immune attack, tumors commonly produce the PD-L1. Furthermore, good patient responses have been linked to increased T cell diversity. If we consider the fraction of T cells PD-1/PD-L1 expression, we can have better predictions (Tumeh et al., 2014).

Tumor Mutations: Mutated proteins that highlight tumors and give targets - neoantigens - for the immune system to attack. Tumor mutation burden (TMB) is a biomarker that measures how many nonsynonymous mutations a tumor acquires per mega-base (Mb) of DNA. The higher number of mutations is correlated with higher number of protein mutations and neoantigens which are targetable by immune system. As a result, cancer patients with a high TMB have been found to be substantially more likely to react to checkpoint immunotherapy. TMB levels are also linked to previous immune responses and PD-1/PD-L1 expression. Patients' responses to anti-CTLA-4 checkpoint immunotherapy are also influenced by the tumor's genetic state (Tumeh et al., 2014). Some cancers lose their capacity to repair DNA, resulting in highly mutated tumors with high microsatellite instability (MSI-H) and deficient mismatch repair (dMMR). MSI-H/dMMR tumors can give better response to ICIs because their tumors are frequently expressing PD-L1 and have infiltrated killer T cells. For example, MSI-H colorectal cancer (CRC) samples have shown significantly stronger responses to immune checkpoint inhibitors (ICIs) (Binnewies et al., 2018). Merck's KEYTRUDA (pembrolizumab), an anti-PD-1 antibody, was authorized in May 2017 as the first medication for patients with MSI-H/dMMR solid tumors (Bai et al., 2020). Some tumor mutations have been linked to primary (Shin et al., 2017) and acquired (Zaretsky et al., 2016) immunotherapy resistance. Other mutations, on the other hand, can be used to develop personalized immunotherapies such as vaccines that trigger a patient's cancer-specific tumor-targeting immune responses. These personalized vaccinations are now being tested in several cancer types in clinical studies.

Tumor-Associated Proteins: Up-regulated gene expression of biomarkers can be utilized in immunotherapy to target cancer cells. HER2, a growth-related protein expressed in healthy cells but frequently produced at unusually high levels in cancer cells, is one such example. Another example is Cancer/testis antigen, typically expressed in adult testicular tissue, but tumor cells may stimulate Cancer/testis antigen synthesis inappropriately. Cancer/testis antigen expression has been linked to more aggressive ovarian cancer (Szender et al., 2017). Furthermore, cells infected with cancer related viruses can produce unusual viral proteins, served as therapeutic targets.

It's doubtful that one ideal biomarker will be relevant in all instances, as the value of many biomarkers will likely vary depending on the kind of cancer and immunotherapy employed. Particular cancer or tumor types may have specific biomarkers. Multiple biomarker panels will be created in the future to give physicians with the most complete and actionable therapies. Discovering and verifying novel biomarkers is still an important topic of study. Technological advancements that allow for deeper investigation will most likely be linked to developments in biomarker discovery. New approaches may potentially make it possible to obtain biomarker information in a less intrusive manner than old procedures. Clinical trials will also be critical in this advancement because they give clinicians a setting to improve immunotherapy effectiveness.

The Cancer Genome Atlas (TCGA) Pan-Cancer project provides data sets of thousands of cancer patients, including simple nucleotide variations (SNVs), gene expression, miRNA expression, DNA methylation, copy number variations (CNVs), clinical and biopsy sample data. TCGA includes data for 33 tumor types, including breast invasive carcinoma (BRCA), glioblastoma multiforme (GBM), ovarian serous cystadenocarcinoma (OV), and lung adenocarcinoma (LUAD) (The Cancer Genome Atlas Research Network, 2013). The TCGA database is a valuable gold mine for integrated cancer analysis because it contains different molecular data for many types of cancer. Therefore, the structure of the data sets in the TCGA database has become a standard. Programs and pipelines created using TCGA data can also be used for databases containing other similar patient data. Because the TCGA project supplies both molecular data and clinical data including drug responses and survival data, these data have been used for many studies including biomarker detection, drug response, and recently machine learning approaches.

In the concept of this project, firstly Thorsson et al., performed a pan-cancer immunologic analysis of 33 TCGA cancer projects and they identified six immunological subgroups based on changes in intra-tumoral heterogeneity, leukocyte fractions, the Th1:Th2 cell ratio, neoantigen load, immunomodulatory gene expression, and tumor prognosis. Specific driver mutations were linked to lower (IDH1, NRAS or CTNNB1) or higher (TP53, BRAF, or CASP8) leukocyte numbers across all cancers (Thorsson et al., 2018). Recently some studies have been conducted to detect prognostic indicators based on the tumor microenvironment such as in 20 cancer types showing poor prognosis (Liu et al., 2020), prostate cancer (Zhu et al., 2020), colorectal cancer (Grasso et al., 2018; Miao et al., 2020, Liao et al., 2021), cervical squamous cell carcinoma (Zhao et al., 2020), head and neck squamous cell carcinoma (Yao et al., 2020), early-stage lung adenocarcinoma (Bao et al., 2020), triple-negative breast cancer (Cheng et al., 2020), ovarian cancer (Cao et al., 2021) and osteosarcoma (Yang et al., 2021). Recently, studies on predictive biomarkers for immunotherapy response have been performed in metastatic urothelial carcinoma (Goswami et al., 2020), acute myeloid leukemia (Vadakekolathu et al., 2020), gastric cancer (Ren et al., 2020), muscle-invasive bladder cancer (Cao et al., 2021) and machine learning methods have been integrated to analyze immunotherapy response in pan-cancer (Xie et al., 2020), non-small cell lung cancer (Wiesweg et al., 2020), lung adenocarcinoma (Peng et al., 2020) by using RNA expression data. These and previous studies mostly concentrated on the detection of predictive or prognostic biomarkers such as PD-1/PD-L1 expression, tumor infiltrated immune cells such as CD8+ T cells, tumor mutation burden (TMB), MSI-H/dMMR status, and individual mutations correlated with each other with taken immunotherapy drugs or not, and lastly for immunotherapy response by using RNA expression data.

Among the indexed clinical data presented for 33 cancers in the TCGA project, 35% (3,786) of the total number of patients (10,690) have drug information. Of 3786 patients with drug information, 1197 of them have a total of 2572 drug response data. There are 16 different antibody therapeutics used for a total of 190 patients and only 99 patients have response data for 8 antibody drugs. We aimed to use TCGA data to cluster patients separately for cancer types and analyze the clusters for clinical and immunological features to detect the predictive or prognostic biomarkers of antibody therapeutics response. To date, approved antibody therapeutics which are used in clinic are summarized in Table 1.1.

Table 1.1. Approved antibody therapeutics used for cancer in clinic.

Therapeutic	Target Protein	Gene Name	Conditions Approved
Aflibercept	Vascular endothelial growth factor (VEGF)	VEGF	Colorectal cancer
Alemtuzumab	CAMPATH-1 antigen (CD52)	CD52	B-cell chronic lymphocytic leukemia (B-CLL)
Amivantamab	Epidermal growth factor receptor (EGFR), Hepatocyte growth factor receptor (HGFR)	EGFR, MET	NSCLC with EGFR exon 20 insertion mutations, Liver cancer
Atezolizumab	Programmed cell death 1 ligand 1 (PD-L1)	CD274	Breast cancer, Liver cancer, Non-small cell lung cancer, Small cell lung cancer, Urogenital cancer, Malignant melanoma
Avelumab	Programmed cell death 1 ligand 1 (PD-L1)	CD274	Merkel cell carcinoma, Urogenital cancer, Renal cell carcinoma
Belantamab mafodotin	Tumor necrosis factor receptor superfamily member 17 (TNFRSF17)	TNFRSF17	Multiple myeloma
Bevacizumab	Vascular endothelial growth factor A (VEGFA)	VEGFA	Breast cancer, Cervical cancer, Colorectal cancer, Glioblastoma, Glioma, Liver cancer, Non-small cell lung cancer, Ovarian cancer, Renal cell carcinoma, Fallopian tube cancer, Peritoneal cancer
Bintrafusp	Programmed cell death 1 ligand 1 (PD-L1)	CD274	Merkel cell carcinoma, Urogenital cancer
Blinatumomab	B-lymphocyte antigen CD19 (CD19), T-cell surface glycoprotein CD3 delta chain (CD3E)	CD19, CD3D	Precursor B-cell lymphoblastic leukaemia-lymphoma
Brentuximab vedotin	Tumor necrosis factor receptor superfamily member 8 (CD30)	TNFRSF8	Hodgkin's lymphoma, T-cell lymphoma, Peripheral T-cell lymphoma, Cutaneous T-cell lymphoma, Anaplastic large cell lymphoma, Primary cutaneous anaplastic large cell lymphoma
Cadonilimab	Programmed Cell Death 1 (PD-1), Cytotoxic T-lymphocyte antigen 4 (CTLA4)	PDCD1, CTLA4	Cervical cancer
Camrelizumab	Programmed cell death protein 1 (PD-1)	PDCD1	Esophageal cancer, Non-small cell lung cancer, Hodgkin's lymphoma, Nasopharyngeal cancer, Liver cancer
Catumaxomab	Epithelial cell adhesion molecule (EpCAM), T-cell surface glycoprotein CD3 epsilon chain (CD3E)	EPCAM, CD3E	Ovarian cancer, Malignant ascites, Gastric cancer

Table 1.1. (continue)

Therapeutic	**Target Protein**	**Gene Name**	**Conditions Approved**
Cemiplimab	Programmed cell death protein 1 (PD-1)	PDCD1	Basal cell carcinoma, Non-small cell lung cancer, Squamous cell carcinoma
Cetuximab	Epidermal growth factor receptor (EGFR)	EGFR	Colorectal cancer, Head and neck cancer
Cetuximab sarotalocan	Epidermal growth factor receptor (EGFR)	EGFR	Head and neck cancer
Daratumumab	ADP-ribosyl cyclase/cyclic ADP-ribose hydrolase 1 (ADPRC1/cADPR hydrolase 1)	CD38	Multiple myeloma
Denosumab	Receptor activator of nuclear factor kappaB ligand (RANKL)	TNFSF11	Bone metastases, Bone cancer
Derlotuximab biotin - I131	DNA/Histone H1 Complex		Lung cancer
Dinutuximab	Ganglioside GD2 (GD2)		Neuroblastoma
Disitamab vedotin	Human epidermal growth factor receptor 2 (HER2)	ERBB2	Gastric cancer, Urothelial carcinoma
Dostarlimab	Programmed cell death protein 1 (PD-1)	PDCD1	Endometrial cancer, Solid tumors
Durvalumab	Programmed cell death 1 ligand 1 (PD-L1)	CD274	Non-small cell lung cancer, Urogenital cancer, Small cell lung cancer
Edrecolomab	Epithelial cell adhesion molecule (EpCAM)	EPCAM	Colorectal cancer
Elotuzumab	SLAM family member 7 (SLAMF7)	SLAMF7	Multiple myeloma
Enfortumab vedotin	Nectin Cell Adhesion Molecule 4 (Nectin 4)	NECTIN4	Urogenital cancer
Envafolimab	Programmed cell death 1 ligand 1 (PD-L1)	CD274	Metastatic microsatellite instability-high (MSI-H) or deficient mismatch repair (dMMR) advanced solid tumors
Gemtuzumab ozogamicin	Myeloid cell surface antigen CD33 (CD33)	CD33	Acute myelogenous leukemia
Glofitamab	B-lymphocyte antigen CD20 (CD20), T-cell surface glycoprotein CD3E (CD3e)	MS4A1, CD3E	Diffuse large B-cell lymphoma
Ibritumomab tiuxetan	B-lymphocyte antigen CD20 (CD20)	MS4A1	Non-Hodgkin's lymphoma
Inetetamab	Human epidermal growth factor receptor 2 (HER2)	ERBB2	HER2-positive metastatic breast cancer

Table 1.1. (continue)

Therapeutic	Target Protein	Gene Name	Conditions Approved
Inotuzumab ozogamicin	B-cell receptor CD22 (CD22)	CD22	B-cell acute lymphoblastic leukemia
Ipilimumab	Cytotoxic T-lymphocyte antigen 4 (CTLA4)	CTLA4	Liver cancer, Non-small cell lung cancer, Mesothelioma, Malignant melanoma, Renal cell carcinoma
Isatuximab	ADP-ribosyl cyclase/cyclic ADP-ribose hydrolase 1 (ADPRC1/cADPR hydrolase 1)	CD38	Multiple myeloma
Leronlimab	C-C chemokine receptor type 5 (CCR5)	CCR5	Breast cancer
Loncastuximab tesirine	B-lymphocyte antigen CD19 (CD19)	CD19	Diffuse large B-cell lymphoma
Margetuximab	Human epidermal growth factor receptor 2 (HER2)	ERBB2	Breast cancer
Metuximab - I131	Basigin (CD147)	BSG	Liver cancer
Mirvetuximab soravtansine	Folate receptor alpha (FR-alpha)	FOLR1	Ovarian cancer, Fallopian tube cancer, Peritoneal cancer
Mogamulizumab	C–C chemokine receptor type 4 (CCR4)	CCR4	Adult T-cell leukemia/lymphoma, Cutaneous T-cell lymphoma, Peripheral T-cell lymphoma
Mosunetuzumab	B-lymphocyte antigen CD20 (CD20), T-cell surface glycoprotein CD3 epsilon chain (CD3e)	MS4A1, CD3E	Follicular lymphoma
Moxetumomab pasudotox	B-cell receptor CD22 (CD22)	CD22	Hairy cell leukemia
Naxitamab	Ganglioside GD2 (GD2)		Neuroblastoma
Necitumumab	Epidermal growth factor receptor (EGFR)	EGFR	Non-small cell lung cancer
Nimotuzumab	Epidermal growth factor receptor (EGFR)	EGFR	Anaplastic astrocytoma, Brain cancer, Esophageal cancer, Glioblastoma, Head and neck cancer, Nasopharyngeal cancer, Glioma
Nivolumab	Programmed cell death protein 1 (PD-1)	PDCD1	Colorectal cancer, Esophageal cancer, Gastric cancer, Head and neck cancer, Non-small cell lung cancer, Hodgkin's lymphoma, Mesothelioma, Squamous cell carcinoma, Urogenital cancer, Malignant melanoma, Renal cell carcinoma

Table 1.1. (continue)

Therapeutic	Target Protein	Gene Name	Conditions Approved
Obinutuzumab	B-lymphocyte antigen CD20 (CD20)	MS4A1	Chronic lymphocytic leukemia, Follicular lymphoma, Non-Hodgkin’s lymphoma
Ofatumumab	B-lymphocyte antigen CD20 (CD20)	MS4A1	Chronic lymphocytic leukemia
Olaratumab	Platelet-derived growth factor receptor alpha (PDGFRA)	PDGFRA	Soft tissue sarcoma
Panitumumab	Epidermal growth factor receptor (EGFR)	EGFR	Colorectal cancer
Pembrolizumab	Programmed cell death protein 1 (PD-1)	PDCD1	Breast cancer, Cervical cancer, Colorectal cancer, Esophageal cancer, Gastric cancer, Head and neck cancer, Liver cancer, Non-small cell lung cancer, Diffuse large B-cell lymphoma, Hodgkin’s lymphoma, Pancreatic cancer, Squamous cell carcinoma, Urogenital cancer, Malignant melanoma
Penpulimab	Programmed cell death protein 1 (PD-1)	PDCD1	Non-small cell lung cancer, Hodgkin’s lymphoma, Nasopharyngeal cancer
Pertuzumab	Human epidermal growth factor receptor 2 (HER2)	ERBB2	Breast cancer
Polatuzumab vedotin	B-cell antigen receptor complex-associated protein beta chain (CD79b)	CD79B	Diffuse large B-cell lymphoma
Prolgolimab	Programmed cell death protein 1 (PD-1)	PDCD1	Melanoma
Pucotenlimab	Programmed cell death protein 1 (PD-1)	PDCD1	Metastatic microsatellite instability-high (MSI-H) or deficient mismatch repair (dMMR) advanced solid tumors,
Racotumomab	Ganglioside GM3 (GM3)		Non-small cell lung cancer
Ramucirumab	Vascular endothelial growth factor receptor 2 (VEGFR-2)	KDR	Colorectal cancer, Gastric cancer, Liver cancer, Non-small cell lung cancer
Relatlimab	Lymphocyte activation gene 3 protein (CD223)	LAG3	Malignant melanoma
Retifanlimab	Programmed cell death protein 1 (PD-1)	PDCD1	Merkel cell carcinoma
Ripertamab	B-lymphocyte antigen CD20 (CD20)	MS4A1	Non-Hodgkin's lymphoma
Rituximab	B-lymphocyte antigen CD20 (CD20)	MS4A1	Chronic lymphocytic leukemia, Diffuse large B cell lymphoma, Follicular lymphoma, Non-Hodgkin's lymphoma

Table 1.1. (continue)

Therapeutic	Target Protein	Gene Name	Conditions Approved
Sacituzumab govitecan	Tumor-associated calcium signal transducer 2 (TACSTD-2)	TACSTD2	Breast cancer, Urogenital cancer
Serplulimab	Programmed cell death protein 1 (PD-1)	PDCD1	Non-small cell lung cancer, Small cell lung cancer, Microsatellite instability-high solid tumors
Sintilimab	Programmed cell death protein 1 (PD-1)	PDCD1	Gastric cancer, Esophageal cancer, Liver cancer, Non-small cell lung cancer, Hodgkin's lymphoma
Sugemalimab	Programmed cell death 1 ligand 1 (PD-L1)	CD274	Non-small cell lung cancer
Tafasitamab	B-lymphocyte antigen CD19 (CD19)	CD19	Diffuse large B-cell lymphoma
Tebentafusp	Melanocyte protein PMEL (GP100), T-cell surface glycoprotein CD3 (CD3)	PMEL, CD3	Uveal melanoma
Teclistamab	Tumor necrosis factor receptor superfamily member 17 (BCMA), T-cell surface glycoprotein CD3 (CD3)	TNFRSF17, CD3	Multiple myeloma
Tislelizumab	Programmed cell death protein 1 (PD-1)	PDCD1	Liver cancer, Non-small cell lung cancer, Hodgkin's lymphoma, Nasopharyngeal cancer, Squamous cell carcinoma, Urogenital cancer
Tisotumab vedotin	Tissue factor (TF)	F3	Cervical cancer
Toripalimab	Programmed cell death protein 1 (PD-1)	PDCD1	Nasopharyngeal cancer, Urogenital cancer, Malignant melanoma
Tositumomab - I131	B-lymphocyte antigen CD20 (CD20)	MS4A1	Non-Hodgkin's lymphoma
Trastuzumab	Human epidermal growth factor receptor 2 (HER2)	ERBB2	Breast cancer, Gastric cancer
Trastuzumab deruxtecan	Human epidermal growth factor receptor 2 (HER2)	ERBB2	Breast cancer, Gastric cancer
Trastuzumab emtansine	Human epidermal growth factor receptor 2 (HER2)	ERBB2	Breast cancer
Tremelimumab	Cytotoxic T-lymphocyte protein 4 (CD152)	CTLA4	Non-small cell lung cancer, Liver cancer
Zimberelimab	Programmed cell death protein 1 (PD-1)	PDCD1	Hodgkin's lymphoma

Antibody therapeutics which are given to patients in TCGA projects are summarized in Table 1.2. As noticed, only eight antibody therapeutics were given to patients in TCGA projects. While most of the approved antibody therapeutics are against PD-1 and PD-L1, patients in TCGA projects have taken only nivolumab which is anti-PD-1 therapeutic and most of the patients have taken bevacizumab which is anti-VEGFA. From Table 1.1 and Table 1.2, the next table were created for possible targetable and targeted proteins in TCGA projects (Table 1.3). We can notice that VEGFA was targeted in many cancer types although bevacizumab is not approved for those particular cancer types (Table 1.3). The patients who took bevacizumab might be in clinical trials during the project. There are also cancer types for which no antibody medication was used (Table 1.3). We aimed to analyze both targetable and targeted proteins in cancer type of interest if it is available.

Table 1.2. Antibody therapeutics given to patients in TCGA projects.

Therapeutic	**Target Protein**	**Gene Name**	**TCGA Projects**
Bevacizumab	Vascular endothelial growth factor A (VEGFA)	VEGFA	ACC, BLCA, BRCA, CESC, COAD, GBM, HNSC, KICH, KIRC, KIRP, LGG, MESO, OV, READ, SARC, SKCM, UCEC, UCS
Cetuximab	Epidermal growth factor receptor (EGFR)	EGFR	CESC, COAD, HNSC, LUSC, OV, STAD
Denosumab	Receptor activator of nuclear factor kappaB ligand (RANKL)	TNFSF11	BRCA
Nivolumab	Programmed cell death protein 1 (PD-1)	PDCD1	BLCA, SKCM
Panitumumab	Epidermal growth factor receptor (EGFR)	EGFR	COAD, HNSC
Pembrolizumab	Programmed cell death protein 1 (PD-1)	PDCD1	LGG, SKCM
Rituximab	B-lymphocyte antigen CD20 (CD20)	MS4A1	BRCA, KIRP
Trastuzumab	Human epidermal growth factor receptor 2 (HER2)	ERBB2	BRCA, OV, SKCM

Table 1.3. Targetable and targeted antigens in TCGA projects.

Study Abbreviation	Study Name	Genes of Targetable Proteins	Genes of Targeted Proteins
ACC	Adrenocortical carcinoma	CD274, PDCD1	VEGFA
BLCA	Bladder urothelial carcinoma	CD274, PDCD1, ERBB2, NECTIN4, TACSTD2	VEGFA, PDCD1
BRCA	Breast invasive carcinoma	CD274, PDCD1, CCR5, ERBB2, TACSTD2, VEGFA	ERBB2, MS4A1, TNFSF11, VEGFA
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	CD274, PDCD1, CTLA4, F3, VEGFA	EGFR, VEGFA
CHOL	Cholangiocarcinoma	CD274, PDCD1	
COAD	Colon adenocarcinoma	CD274, PDCD1, EGFR, EPCAM, KDR, VEGFA	EGFR, VEGFA
DLBC	Lymphoid neoplasm diffuse large B-cell lymphoma	CD19, CD79B, MS4A1, CD3E, PDCD1	
ESCA	Esophageal carcinoma	CD274, PDCD1, EGFR	
GBM	Glioblastoma multiforme	CD274, PDCD1, EGFR, VEGFA	VEGFA
HNSC	Head and neck squamous cell carcinoma	CD274, PDCD1, EGFR	EGFR, VEGFA
KICH	Kidney chromophobe	CD274, PDCD1, ERBB2, NECTIN4, TACSTD2	VEGFA
KIRC	Kidney renal clear cell carcinoma	CD274, PDCD1, CTLA4, ERBB2, NECTIN4, TACSTD2, VEGFA	VEGFA
KIRP	Kidney renal papillary cell carcinoma	CD274, PDCD1, CTLA4, ERBB2, NECTIN4, TACSTD2, VEGFA	MS4A1, VEGFA
LAML	Acute myeloid leukemia	CD33	
LGG	Brain lower grade glioma	CD274, PDCD1, EGFR, VEGFA	VEGFA, PDCD1
LIHC	Liver hepatocellular carcinoma	CD274, PDCD1, BSG, CTLA4, MET, KDR, VEGFA	
LUAD	Lung adenocarcinoma	CD274, PDCD1, CTLA4, EGFR, MET, KDR, VEGFA	
LUSC	Lung squamous cell carcinoma	CD274, PDCD1, CTLA4, EGFR, MET, KDR, VEGFA	EGFR
MESO	Mesothelioma	CD274, PDCD1, CTLA4	VEGFA
OV	Ovarian serous cystadenocarcinoma	CD274, PDCD1, EPCAM, CD3E, FOLR1, VEGFA	ERBB2, EGFR, VEGFA

Table 1.3. (continue)

Study Abbreviation	**Study Name**	**Genes of Targetable Proteins**	**Genes of Targeted Proteins**
PAAD	Pancreatic adenocarcinoma	CD274, PDCD1	
PCPG	Pheochromocytoma and paraganglioma	CD274, PDCD1	
PRAD	Prostate adenocarcinoma	CD274, PDCD1, ERBB2, NECTIN4, TACSTD2	
READ	Rectum adenocarcinoma	CD274, PDCD1	VEGFA
SARC	Sarcoma	CD274, PDCD1, PDGFRA	VEGFA
SKCM	Skin cutaneous melanoma	CD274, PDCD1, CTLA4, EPCAM, CD3E, LAG3	ERBB2, VEGFA, PDCD1
STAD	Stomach adenocarcinoma	CD274, PDCD1, EPCAM, CD3E, ERBB2, KDR	EGFR
TGCT	Testicular germ cell tumors	CD274, PDCD1, ERBB2, NECTIN4, TACSTD2	
THCA	Thyroid carcinoma	CD274, PDCD1	
THYM	Thymoma	CD274, PDCD1	
UCEC	Uterine corpus endometrial carcinoma	CD274, PDCD1	VEGFA
UCS	Uterine carcinosarcoma	CD274, PDCD1	VEGFA
UVM	Uveal melanoma	CD274, PDCD1, PMEL, CD3	

2. CHAPTER

Comprehensive Profiling of Genomic and Transcriptomic Differences Between Risk Groups of Lung Adenocarcinoma and Lung Squamous Cell Carcinoma[1]

2.1. Abstract

We analyzed genomic and transcriptomic variations in subtypes of non-small-cell lung cancer, lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC). The goal was to identify key genes for prognostic prediction in these cancer types. We used The Cancer Genome Atlas (TCGA) gene expression data to develop gene signatures for prognostic risk prediction. We generated 35-gene and 33-gene signatures for LUAD and LUSC, respectively, and found that clustering patients into high- and low-risk groups based on these signatures showed significant results with high prediction power. We also identified specific genomic alterations related to risk groups or cancer types, including differences in immune and metabolic pathways. High-risk groups of both LUAD and LUSC have more downregulated immune pathways with upregulated metabolic pathways. Several important gene alterations were observed, such as deletions in CDKN2A and CDKN2B, SOX2 amplification in LUSC, and PSMD4 amplification in LUAD. Mutations in EGFR, KRAS, and other genes were found to be subtype specific. The low-risk groups had a higher number of genomic alterations. We suggest that the identified signature genes and altered genes could potentially be used as prognostic biomarkers in personalized oncology.

---

[1] **Comprehensive Profiling of Genomic and Transcriptomic Differences between Risk Groups of Lung Adenocarcinoma and Lung Squamous Cell Carcinoma**
T. Zengin and T. Önal-Süzek. *Journal of Personalized Medicine, 2021, doi: 10.3390/jpm11020154*

2.2. Methods for Risk Clustering

Simple nucleotide variations (Masked Somatic Mutations generated by mutect2 pipeline), transcriptome profiling/RNA expression (RNAseq - HTSeq counts), copy number variations (masked copy number segment) and clinical data of patients in the LUAD and LUSC projects were downloaded from harmonized database (data aligned against hg38 genome) of NCI GDC (Grossman et al., 2016) using the TCGAbiolinks R package (Colaprico et al., 2016). All patients do not have all types of data, therefore we used the subset of the patients who have these four types of data.

Gene expression quantification data (HTSeq counts) of patients with unpaired RNAseq data (tumor samples without corresponding normal samples) were processed. First, the data was normalized using the TMM method, followed by a log2 transformation. Genes with consistently zero or low counts were filtered out. Next, univariate cox proportional hazards regression analysis was performed to identify genes associated with patient survival. Among these survival-related genes ($p \leq 0.05$), a lasso-regularized cox model was applied using the glmnet R package (Simon et al., 2011). The model's performance was evaluated using leave-one-out cross-validation (LOOCV). Subsequently, multivariate cox regression was performed for the signature genes, and the risk score for each patient was calculated based on the multivariate cox regression model. Patients were then classified into high-risk and low-risk groups based on determined risk score threshold, the best cutoff value for ROC calculated by cutoff R package (Zhang and Jin, 2019). To validate the gene signature, HTSeq counts from tumor samples with paired RNAseq data (tumor samples along with adjacent normal samples) were used. Similar preprocessing steps were applied, including data filtering, TMM normalization, and log2 transformation. Multivariate cox regression was performed for the signature genes, and the predictive performance of the model was evaluated. The risk score for each patient in the validation group was calculated using the multivariate cox regression model, and patients were assigned to either the high-risk or low-risk group.

Differentially expressed genes ($q \leq 0.01$) were identified using the limma (Ritchie et al., 2015) and edgeR (McCarthy et al., 2012) R packages. Active subnetworks of these differentially expressed genes were determined using the DEsubs R package (Vrahatis et al., 2016). Furthermore, significantly aberrant copy number variations in tumor samples were identified and visualized using the gaia R package (Morganella et al., 2020). Somatic mutations were filtered and categorized as either oncogenes (OG) or tumor suppressor genes (TSG) using the SomInaClust R package (Eynden et al., 2015). To identify hotspots, SomInaClust utilized a background mutation value based on known somatic mutations from the "COSMIC" and "Cancer Gene Census" (v92) datasets of the COSMIC database (Tate et al., 2018). These analyses were performed separately for high-risk and low-risk groups of LUAD and LUSC patients. The overall methodology is summarized in Figure 2.1 as a flowchart.

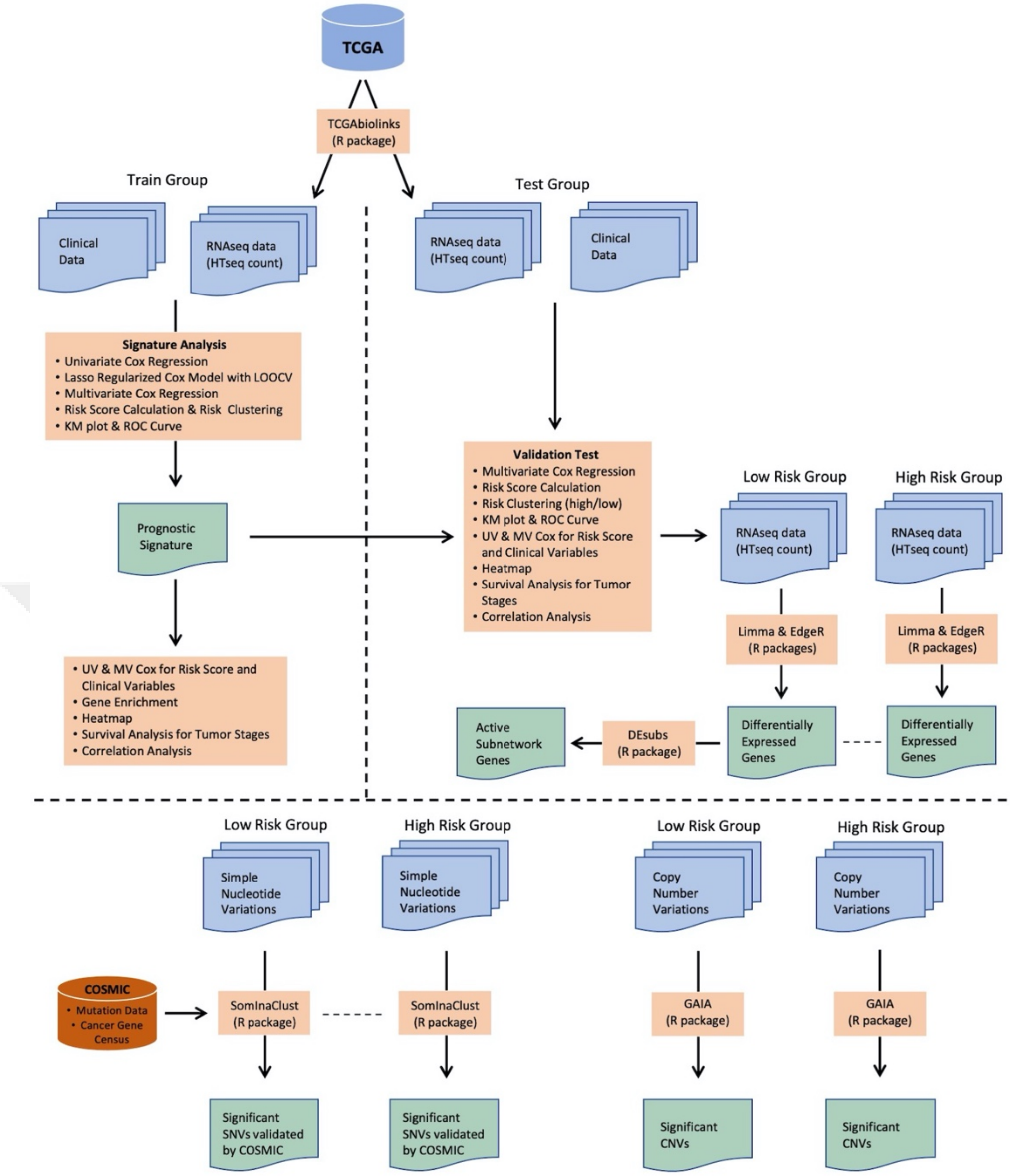


Figure 2.1. Flowchart of the methods used for risk clustering and analysis of genomic and transcriptomic variations between risk groups.

2.3. Results of Risk Clustering

In order to validate the gene expression signature, we predicted the risk scores of each patient in the LUAD test group and classified them into high-risk and low-risk groups based on the risk scores. The risk groups exhibited distinct patterns of gene expression of the signature genes, with the high-risk group showing shorter survival times and a higher number of deaths, resulting in a significantly different survival probability (p <

0.0001). The risk score demonstrated strong predictive power, with AUC values of 0.97, 0.92, 0.93, and 0.92 for 1, 3, 5, and 8 years, respectively (Figure 2.2).

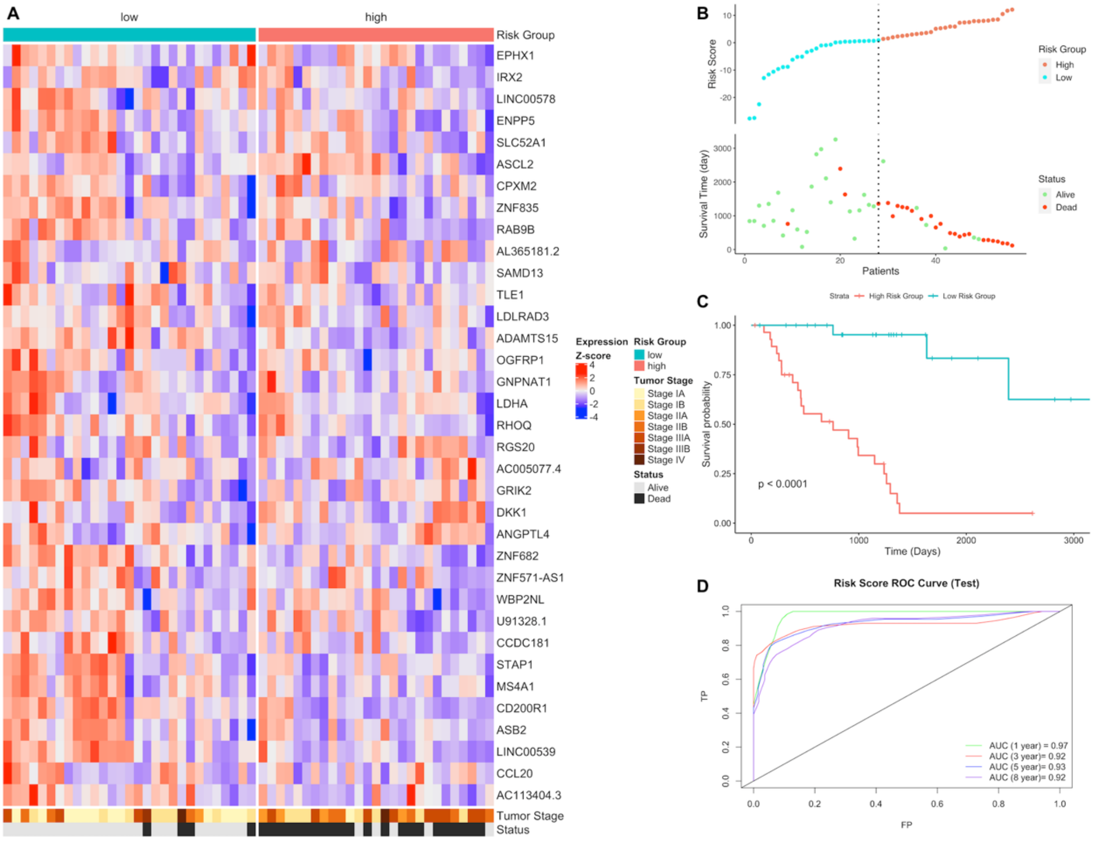


Figure 2.2. Gene expression signature and risk clustering of LUAD test dataset. Test dataset patients were clustered into high- and low-risk groups based on risk scores calculated using signature genes expression. (A) Expression heatmap of the signature genes in tumor samples of LUAD patients. (B) Scatter plot showing risk scores, survival time and separation point of the patients into risk groups. (C) KM survival plot showing the overall survival probability between risk groups. (D) ROC curve showing prediction power of risk score in the test dataset for 1, 3, 5 and 8 years.

We predicted the risk scores for each patient in the LUSC test group using the gene signature and classified them into high-risk and low-risk groups. The risk groups displayed distinct patterns of gene expression in the signature genes. The risk groups showed significantly different survival probabilities ($p < 0.0001$) with the high-risk group exhibiting shorter survival times and a higher number of deaths. The risk score demonstrated strong predictive power, with AUC values of 0.93, 0.95, 0.96, and 0.97 for 1, 3, 5, and 8 years, respectively (Figure 2.3).

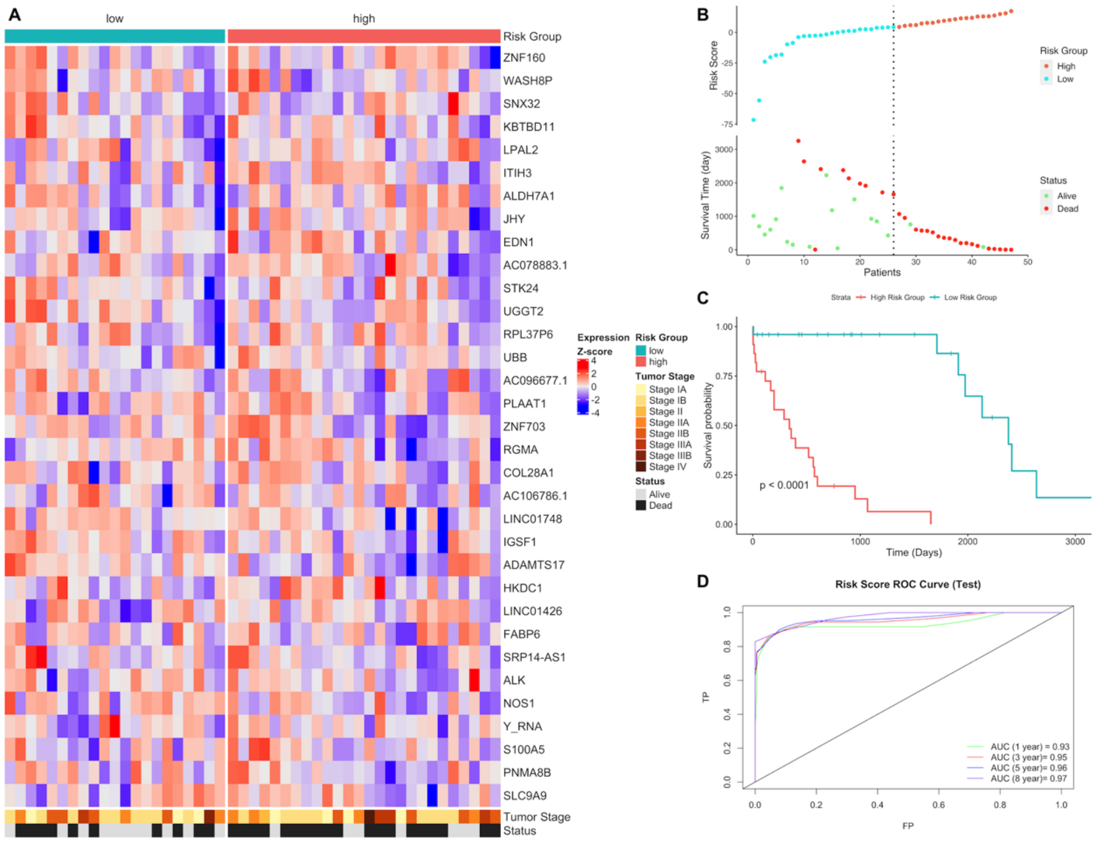


Figure 2.3. Gene expression signature and risk clustering of LUSC test dataset. Patients were clustered into high- and low-risk groups based on risk score. (A) Expression heatmap of the signature genes in tumor samples of LUSC patients. (B) Scatter plot showing risk scores, survival time and separation point of the patients into risk groups. (C) KM survival plot showing the overall survival probability between risk groups. (D) ROC curve showing prediction power of risk score in the test dataset for 1, 3, 5, and 8 years.

Although the LUAD and LUSC expression gene signatures have no genes in common, they share eight pathways, most of which are metabolic pathways including central carbon metabolism, glycolysis/gluconeogenesis, HIF-1 signaling pathway, pyruvate metabolism, PPAR signaling pathway, amino and nucleotide sugar metabolism, TNF signaling pathway, and neurodegenerative pathways in cancer - multiple diseases.

The full version of the article can be found at Appendix A.

2.4. Discussion of the method

The accurate prediction of disease prognosis is crucial in cancer management but remains challenging for many types of cancer because of limitations in pathological and clinical variables. Usage of gene expression signatures have emerged as a promising approach to predict patient's outcome by separating patients into distinct clusters for the guidance of personalized treatment. The first example, 70-gene signature showed prognostic power by categorizing patients with early-stage breast cancer into groups to predict prognosis. It is now commercially available as MammaPrint test which is routinely employed to decide adjuvant chemotherapy. Since then, significant attention has been focused on identifying gene expression-based prognostic signatures for different types of cancer. However, only a limited number of these signatures have been officially approved for clinical use. Typically, most reported prognostic signatures, including those based on functional categories, categorize patients into different risk groups by comparing their risk scores derived from expression levels of signature genes. However, the risk thresholds established using training datasets cannot be directly applied to independent datasets due to inherent biases in measurements caused by batch effects and platform variations causing low reproducibility in independent data (Qian et al., 2021). The classification of patients using signature genes is not commonly employed in clinical practice for many cancer types and faces challenges in terms of standardization. For example, the subtyping of colorectal cancer (CRC) using CMS (Consensus Molecular Subtypes) is not commonly employed in clinic due to several limitations. These limitations include a lack of standardization and the need for bioinformatics resources. These factors restrict the more extensive utilization of CMS subtyping in routine clinical settings (Qian et al., 2021). On the other hand, gene expression panels commercially available to predict prognosis in various diseases are designed to be user-friendly. However, these panels often lack a clear biological interpretation and predictive value. As a result, their usage should be limited to specific clinical scenarios where their utility has been well-established. Another example, currently, none of the biomarkers based on hepatocellular carcinoma (HCC) signatures are routinely used in clinical practice. Several factors contribute to this situation, including challenges related to biology, technology, statistics, and informatics. One primary factor is that gene signatures are commonly developed through retrospective studies, which inherently carry certain biases and are susceptible to confounding factors. These limitations arise due to the nature of retrospective research design, potentially affecting the reliability and generalizability of the developed gene signatures. Although retrospective studies can establish an association between gene signatures and prognosis, their findings often lack reproducibility and cannot be readily translated into clinical applications. The limitations in reproducibility and applicability arise from various factors, including the heterogeneity of study populations, differences in data analysis methods, and potential biases present in retrospective study designs. As a result, further validation and prospective studies are necessary to determine the clinical utility of gene signatures in a more reliable and applicable manner (Qian et al., 2021). Additionally, validation is crucial to assess the predictive accuracy of gene signatures, ideally through independent patient cohorts. As a sad example, Pinyol et al., verified 22 HCC gene signatures published previously, but none of them were able to accurately predict recurrence (Pinyol et al., 2019). Improved gene expression analysis in hepatocellular carcinoma, colorectal cancer, and breast cancer can be achieved through several potential solutions. These include establishing standardization of sample collection and preservation, as well as ensuring the consistent use of diagnostic algorithms and parameters for gene expression classification. The integration of multiple biological levels, such as incorporating transcriptomic data, has shown promise in various publications. The emergence of integrative methods has potential to advance the field and enhance clinical outcomes for these deathly diseases.

3. CHAPTER

Integrative Clustering and Comparison of Clusters for Immune Targetable Cancer Biomarkers

3.1. Introduction

Cancer is characterized by its intricate and diverse nature, making it a complex and heterogeneous because of various genetic alterations, including chromosomal rearrangements, somatic mutations, epigenetic changes, and differential gene expression. While tumors may share similar histopathologic features, their genomic profiles can vary significantly, leading to divergent responses to identical treatments and distinct clinical outcomes. Hence, it is imperative to categorize tumors into molecular subtypes and identify key molecular alterations as drivers, which can be targeted through precision medicine approaches. The objective of cancer genome projects is to unveil critical genetic alterations in cancer and identify potential therapeutic targets by extensively examining the genomic landscape. Prominent initiatives like The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) have been established to carry out extensive studies using advanced techniques such as whole-genome or whole-exome sequencing, profiling of DNA methylation, chromosomal copy number variation, and mRNA/miRNA expression. These efforts have generated an immense volume of data for many types of cancer. The aim is to comprehensively catalog individual genomic alterations and uncover common biological principles. However, the considerable heterogeneity observed in cancer genomes and differences between individuals exhibiting abnormalities in different genes, poses a significant challenge in identifying functionally important genes with therapeutic implications. Therefore, there is a critical need to integrate the vast amount of data types and extract biological principles to advance diagnosis and prognosis (Mo et al., 2013; Mo et al., 2018).

3.1.1. Integrative clustering

A method called iCluster (Shen et al., 2009) offers an innovative approach to integrative clustering by leveraging the relationship between Principal Component Analysis (PCA) and latent variable models. iCluster computes the principal components by employing maximum-likelihood estimation of parameters within a Gaussian latent variable model. This method formulates the K-means as both Gaussian latent variable model and joint latent variable model that incorporates multiple data types. Within a unified framework, it effectively models the associations among various data types and the variance-covariance structure within each data type. Furthermore, iCluster reduces the dimensionality of the datasets, simultaneously. The extraction process relies on the Expectation-Maximization algorithm, utilizing a lasso-type regularization method in the penalized complete-data likelihood. This approach allows for the modeling of tumor subtypes as latent variables, which can be estimated from available data types (Shen et al., 2009). In their study, Shen et al. utilized a dataset containing DNA copy number and RNA expression data obtained from cDNA microarrays of 6691 genes for integrative clustering of 37 breast cancer samples and 4 breast cancer cell lines, resulted in 4 clusters. They also analyzed 91 lung adenocarcinomas samples containing copy number and gene expression of 2782 genes and found 4 clusters of samples (Shen et al., 2009). Later, Shen et al. used the iCluster method for DNA copy number, DNA methylation and RNA expression of 55 glioblastoma (GBM) samples separated in 3 clusters which have distinct clinical features (Shen et al., 2012).

iClusterPlus (Mo et al., 2013) is the improved version of the method integrating more diverse data types: binary data (DNA mutation: 0, 1), categorical data (DNA copy number: amplification, normal, deletion), and continuous data (mRNA expression, miRNA expression, DNA methylation) (Figure 3.1). They analyzed 189 colorectal carcinoma (CRC) samples that had all four data types (DNA somatic mutation, DNA copy number, DNA methylation, and RNA expression) resulting in 4 clusters.

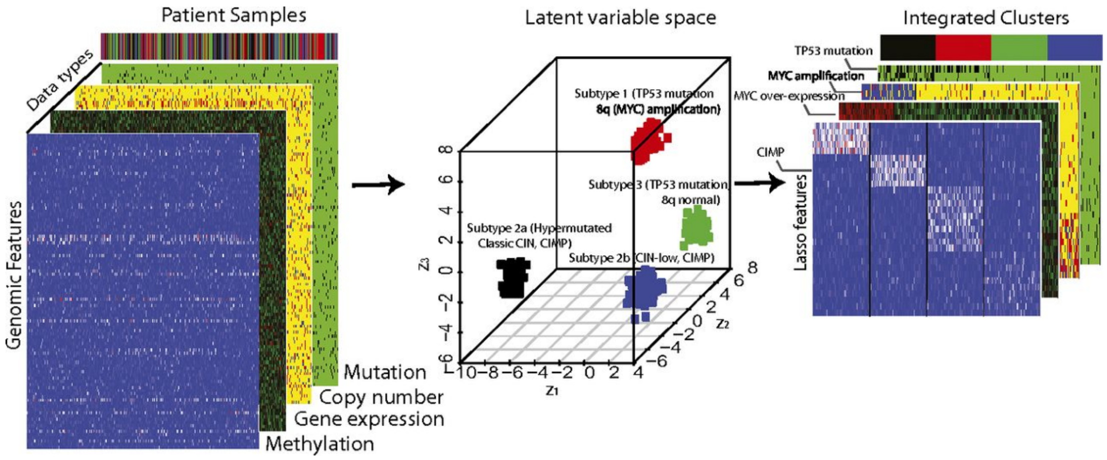


Figure 3.1. Integration of diverse data types by a latent variable approach. (Taken from Mo et al 2013)

The iClusterPlus algorithm has been applied in many cancer genomics studies, encompassing various types of cancers such as lung squamous cell carcinoma (LUSC), lung adenocarcinoma (LUAD), stomach adenocarcinoma (STAD) and uterine corpus endometrial carcinoma (UCEC) (TCGARN, 2012; 2013; 2014a; 2014b). However, it has some limitations and therefore, they developed iClusterBayes method (Mo et al., 2018) which performs Bayesian integrative clustering (Figure 3.2) that overcomes the limitations of the previous method. iClusterBayes significantly improved statistical computation and calculates a posterior probability for each data feature, which is a great advantage of this new method.

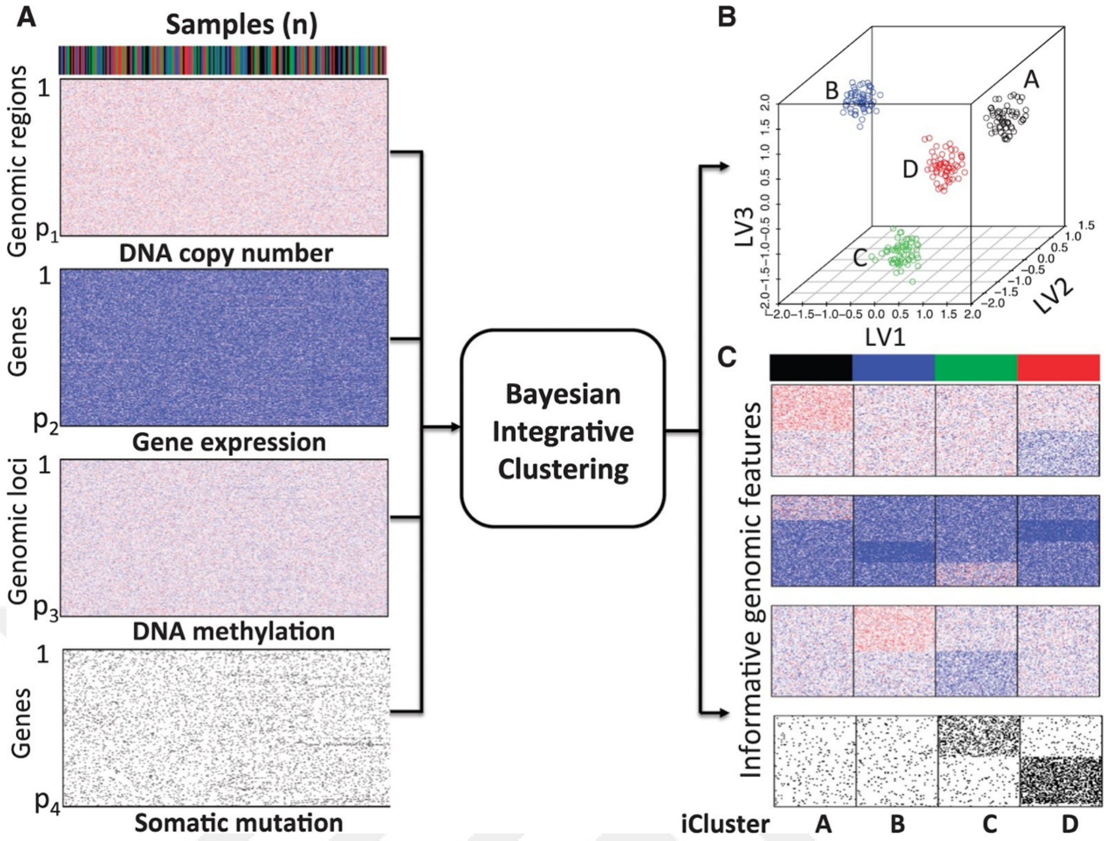


Figure 3.2. The proposed Bayesian integrative clustering framework. (A) Multiple genomic data sets are modelled jointly by bayesian latent variable regression models. (B) Detect common latent variables from previous step are used in lower dimensional integrated latent variable space to cluster patient samples. (C) Driver data features contributing to sample clustering are determined. (Taken from Mo et al 2018)

Mo et al analyzed 241 kidney renal clear cell carcinoma (KIRC) samples containing SNV, CNV, DNA methylation, RNA expression and miRNA expression data. They used genes with somatic mutation rate >2%, 4470 non-redundant copy number regions, log2-transformed normalized top 20% (4106) most-variable RNA expression, top 20% (3955) most-variable DNA methylation sites and log2 transformed and the top 80% most-variable miRNA data and iClusterBayes method resulted in 2 clusters of patient samples.

They also analyzed 84 glioblastoma (GBM) samples containing somatic mutation, DNA copy number and RNA expression (1740 most variable gene expression) using iClusterBayes which resulted in 4 clusters. The mutated genes which drive subtype clustering are NF1, TP53, MN1 (tumor suppressor genes), NF1, TP53, MAPK9, MAPK7, PIK3R1 (intracellular signaling cascades), and A2M, ITGB2, FN1 (inflammatory and defense responses). From CNV data 161 regions on chromosomes 4, 7, 9, and 12 were the drivers of clustering. These genomic regions encompass significant genes involved in tumor suppression, such as CDKN2A and CDKN2B on chromosome 9. Additionally, they contain oncogenes that play crucial roles in regulating cell activation, division, and proliferation, such as PDGFRA on chromosome 4, EGFR on chromosome 7, and CDK4, MDM2 and TSPAN31 on chromosome 12. In the RNA expression data set, 711 genes were identified as the drivers, 204 of which play a role in nervous system development, while 507 genes are working in immunologic responses, cell proliferation, adhesion and migration.

We used iClusterBayes for this project because it needs less computing resource if the used data is large. Moreover, iClusterBayes calculates posterior probability of genomic features that drive the integrative clustering. We aimed to compare integrative clusters created for 33 cancers in the TCGA project using simple nucleotide variations (SNV), copy number variations (CNV), methylation levels, transcriptome profiling data (Gene expression and miRNA expression) of patients, with clinical characteristics of the patients in these clusters. As eight different antibody therapeutics were given to patients in the TCGA project, the status of targeted proteins was aimed to be analyzed among patient clusters using expression data. At the end, we aimed to integrate integrative cluster information with curated TCGA subtypes to the web tool called TCGAnalyzeR designed by our group.

3.2. Methods

3.2.1. Data

Simple Nucleotide Variations (Masked Somatic Mutations), gene expression (RNAseq - HTSeq counts), miRNA expression (miRNAseq), Copy Number Variations (Masked Copy Number Segment), DNA methylation (Illumina Human Methylation 450 - Methylation Beta Value) and clinical data of 33 TCGA cancer projects were downloaded from harmonized database (data aligned against hg38 genome) of NCI GDC (Grossman et al., 2016) using the TCGAbiolinks R package (Colaprico et al., 2016). Number of patients with tumor samples are summarized by cancer type and data type (Table 3.1).

Table 3.1. Number of patients with tumor samples by cancer types for five different data types.

Project	Project Name	SNV	CNV	Gene Exp	miRNA Exp	DNA Methylation	All Data Types
ACC	Adrenocortical carcinoma	92	90	79	80	80	**77**
BLCA	Bladder urothelial carcinoma	412	412	408	409	412	**405**
BRCA	Breast invasive carcinoma	986	1096	1091	1078	782	**676**
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	289	295	304	307	307	**276**
CHOL	Cholangiocarcinoma	50	36	36	36	36	**35**
COAD	Colon adenocarcinoma	399	452	456	444	296	**289**
DLBC	Lymphoid neoplasm diffuse large B-cell lymphoma	37	48	48	47	48	**37**
ESCA	Esophageal carcinoma	184	184	161	184	185	**160**
GBM	Glioblastoma multiforme	393	596	154	0	140	**47**
HNSC	Head and neck squamous cell carcinoma	508	522	500	523	528	**484**
KICH	Kidney chromophobe	66	66	65	66	66	**65**
KIRC	Kidney renal clear cell carcinoma	336	532	530	516	319	**256**
KIRP	Kidney renal papillary cell carcinoma	281	290	288	291	275	**264**
LAML	Acute myeloid leukemia	143	200	151	188	140	**76**

Table 3.1. (continue)

Project	Project Name	SNV	CNV	Gene Exp	miRNA Exp	DNA Methylation	All Data Types
LGG	Brain lower grade glioma	508	515	511	512	516	**498**
LIHC	Liver hepatocellular carcinoma	364	375	371	372	377	**353**
LUAD	Lung adenocarcinoma	567	518	513	513	458	**447**
LUSC	Lung squamous cell carcinoma	492	503	501	478	370	**359**
MESO	Mesothelioma	82	87	86	87	87	**81**
OV	Ovarian serous cystadenocarcinoma	436	587	374	489	10	**7**
PAAD	Pancreatic adenocarcinoma	178	184	177	178	184	**169**
PCPG	Pheochromocytoma and Paraganglioma	179	178	178	179	179	**177**
PRAD	Prostate adenocarcinoma	495	497	495	494	498	**485**
READ	Rectum adenocarcinoma	137	165	166	161	98	**94**
SARC	Sarcoma	237	260	259	259	261	**233**
SKCM	Skin cutaneous melanoma	467	104	103	97	104	**97**
STAD	Stomach adenocarcinoma	437	442	375	436	395	**332**
TGCT	Testicular germ cell tumors	144	150	150	150	150	**144**
THCA	Thyroid carcinoma	492	505	502	506	507	**484**
THYM	Thymoma	123	124	119	124	124	**118**
UCEC	Uterine corpus endometrial carcinoma	530	540	543	538	431	**407**
UCS	Uterine carcinosarcoma	57	56	56	57	57	**55**
UVM	Uveal melanoma	80	80	80	80	80	**80**
TOTAL		**9514**	**10013**	**9151**	**9204**	**7932**	**7225**

3.2.2. Preparation of data for integrative clustering

Simple nucleotide variations (masked somatic mutations) data as mutation annotation format (maf) files generated by mutect2 pipeline were filtered to subset genes with somatic mutation rate >1% using the maftools (Mayakonda et al., 2018).

Copy number variation (masked copy number segment) data of the primary tumor samples were filtered to extract non-redundant copy number regions. We reduce the copy number regions consisting of multi-sample array data to 1K-5K non-redundant regions by removing the redundant regions.

Gene expression quantification data (HTSeq counts) of the tumor samples were normalized by TMM method and log2 transformed after removing genes that have zero or low counts. The genes which have highly variable expression were filtered by removing genes exhibiting variation less than 90% across samples.

miRNA expression quantification data (miRNAseq data) of the tumor samples were normalized by TMM method and log2 transformed after removing miRNAs which have zero or low counts.

DNA methylation data (Illumina Human Methylation 450 - Methylation Beta Value) were filtered by removing methylated CpG sites exhibiting variation less than 99% across samples to reduce the huge amount of methylation data that is advantageous for downstream integrative clustering analysis.

3.2.3. Integrative clustering analysis

Patients with tumor samples were clustered depending on prepared datasets using iClusterBayes method of iClusterPlus R package (Mo et al., 2018) which performs integrative clustering of patient samples using multiple data types. If the number of clusters in the samples is known, we can select the appropriate value of k (the number of latent variables) before clustering. The general rule is that the number of clusters is equal to k+1. In cases where the cluster number is unknown, we can test different values of k ranging from 1 to N, where N represents a reasonable number of clusters. In this study, the range of k was set from 1 to 10. Using parallel computing, iClusterBayes was run to search for the best model. In order to determine the best sparse model with the optimal combination of penalty parameters, we used bayesian information criteria (BIC) for each k (from 1-10). Additionally, to select the most suitable value for k, we calculated the deviance ratio, which can be interpreted as the percentage of explained variation (deviance ratio). The preferable selection for k would be based on the lower values of BIC and higher values of deviance ratio (explained variation). In cases where the data contains noise, there is a possibility that the deviance ratio will continue to rise, and the BIC will continue to decrease as the value of k increases. In this situation, it is recommended to generate heat maps of the datasets and determine the optimal number of clusters by analyzing the patterns exhibited by the features on these heat maps. We chose the k where the curve of BIC and deviance ratio levels off. The process was run as parallel to 10 cores and just for the GBM dataset, the whole process took 5 CPU hours using 0.8 GB RAM and taking up 626 MB hard disk space for SNV, CNV and RNAseq data of 146 patients.

3.2.4. Feature selection with gene and pathway enrichment

After selection of k which has lower BIC value and highest deviance ratio, significant features for each type of used dataset which have bayesian posterior probability bigger than 0.5 were selected for visualization and enrichment analyses. For gene enrichment, genes on copy number segments were determined with their HUGO symbols and NCBI Gene identifiers using GenomicRanges R package (Lawrence et al., 2013). HUGO symbols and NCBI Gene IDs of the of mRNAseq genes and miRNAs were determined using the biomaRt R package (Durinck et al., 2009). Pathway enrichment from KEGG database was performed for selected genes/miRNAs using clusterProfiler R package (Yu et al., 2012).

3.2.5. Curated TCGA data from publications

We used several subtype information and tumor sample related data were derived from publications based on TCGA data analyses. The values of overall survival (OS), disease-specific survival (DSS), disease-free interval (DFI) and progression-free interval (PFI) were obtained from curated clinical data resource (Liu et al., 2018). Immune subtypes data, homologous DNA repair deficiency (HRD) data derived from (Knijnenburg et al., 2018), Tumor mutation burden (TMB) data, tumor purity data generated using ABSOLUTE (Carter et al., 2012, Taylor et al., 2018), Tumor infiltrated leukocytes (TILs) estimated based on DNA methylation assays, immune cellular fraction estimated using CIBERSORT (Newman et al., 2015) and TCGA subtypes curated from previous TCGA publications were downloaded from the supplemental data of “The Immune Landscape of Cancer” article (Thorsson et al., 2018).

3.2.6. Statistical Analysis

Categorical variable comparison, continuous variable comparison across multiple groups and correlation between numeric variables in the dataset were performed using ggstatsplot R package (Patil et al., 2021). Expression and subtypes heatmaps were drawn by ComplexHeatmap R package (Gu et al., 2016) with color palettes from wesanderson R package. Survival analysis was performed using survival (Therneau et al., 2020) and survminer (Kassambara et al., 2020) R packages.

3.2.7. Analysis of the known targets and the approved/used immunotherapeutics

The list of approved antibody therapeutics and target proteins of these antibodies were obtained from Thera-SAbDab database (Raybould et al., 2020). The list of used antibody therapeutics with response information in clinical data of TCGA data were derived using in-house R scripts. The expression levels of proteins targeted by the possible and used antibody therapeutics were examined by heatmaps and violin plots using ComplexHeatmap and ggstatsplot R packages (Gu et al., 2016; Patil et al., 2021).

3.2.8. Integrating the iCluster data into the web tool

All curated subtype and iCluster data were integrated into the R-shiny based web tool called TCGAnalyzeR (tcganalyzer.mu.edu.tr) designed by our research group. This web tool enables data visualization and survival analysis for users.

All methods used for this project are summarized in Figure 3.3.

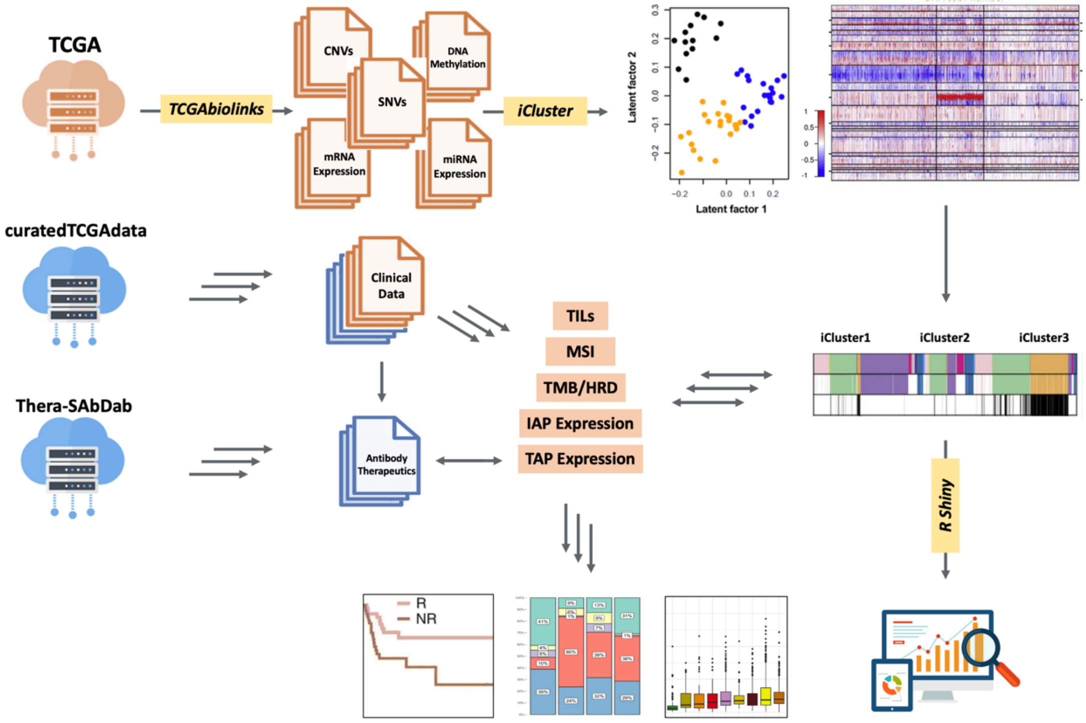


Figure 3.3. The flowchart summarizing the methods used for integrative clustering and comparison of clusters for immune targetable cancer biomarkers. CNV: Copy Number Variation, TIL: Tumor Infiltrated Leukocytes, MSI: Micro-Satellite Instability, TMB: Tumor Mutation Burden, HRD: Homologous DNA Repair Deficiency, IAP: Immune Associated Protein, TAP: Tumor Associated Protein.

3.3. Results

3.3.1. Integrative clustering of LUAD patients

We clustered 447 Lung adenocarcinoma (LUAD) tumor samples using SNV, CNV, gene expression, miRNA expression and methylation data. We used genes with somatic mutation rate >1%, non-redundant copy number regions, log2-transformed normalized top %10 most-variable gene expression data, all miRNA expression after low count filter, and top 1% most-variable DNA methylation after NA removing. From the BIC and deviance ratio plots, we saw that k=3 (k+1=4 clusters) is the most optimal solution (Figure 3.4).

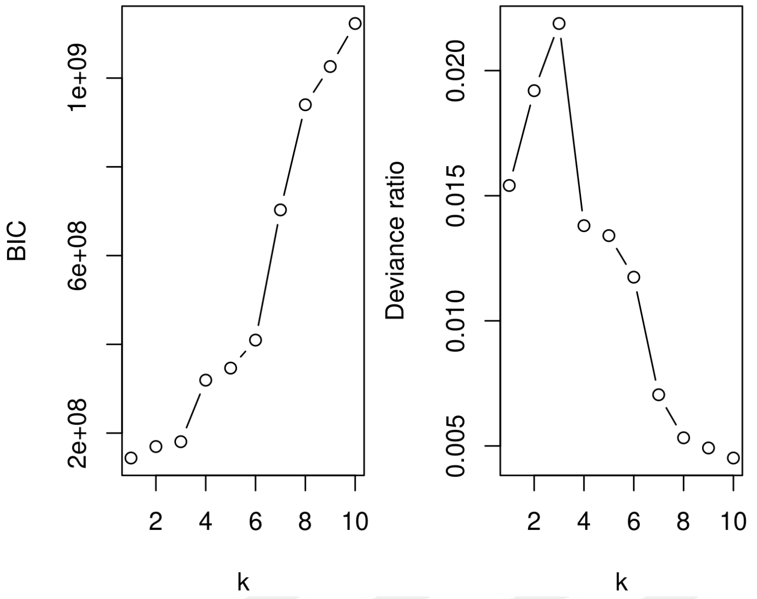


Figure 3.4. Bayesian Information Criteria (BIC) and the deviance ratio vs k for LUAD dataset.

iClusterBayes analysis showed that the number of genes containing SNVs with posterior probabilities is 1225, while the number of expressed genes is 1788 and the number of expressed miRNAs is 198, which have posterior probabilities >0.5. However, there are no CNV segments and methylation sites which have posterior probabilities >0.5, therefore CNV and methylation data are not informative for clustering for this dataset (Figure 3.5). Gene expression is highly dominant to drive integrative clustering followed by SNVs and miRNA expression of LUAD samples (Figure 3.5).

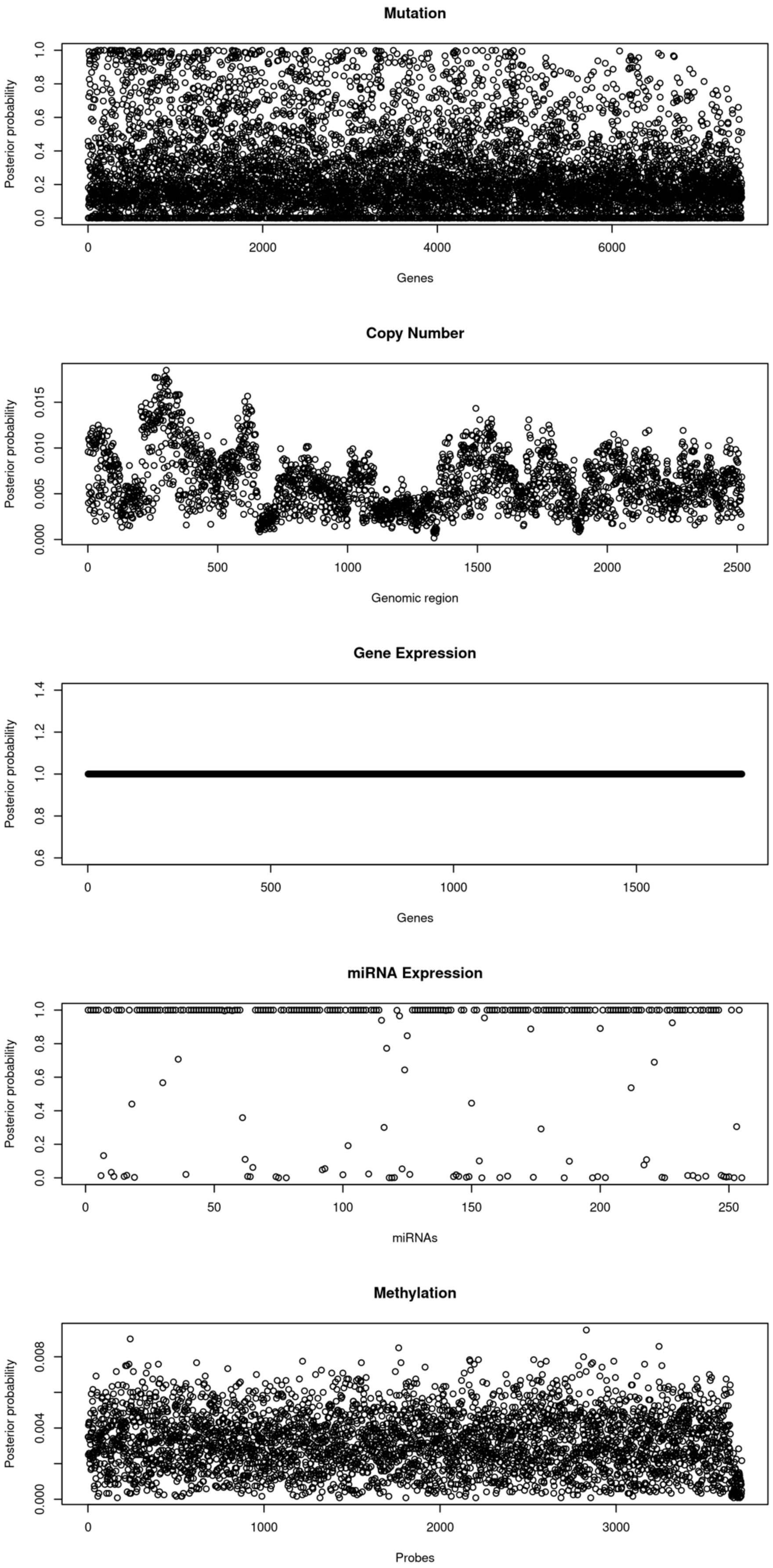


Figure 3.5. Posterior probabilities of genomic features that drive the integrative clustering.

According to Figure 3.6, we can observe significant differences between clusters in heatmaps of SNV, gene expression and miRNA expression. For SNVs, iCluster2 has least mutations while iCluster2 has highest number of mutations. For gene expression, iCluster2 has highest expression while iCluster1 has lowest expression of same set of genes. On the other hand, miRNA expression patterns have differences between clusters.

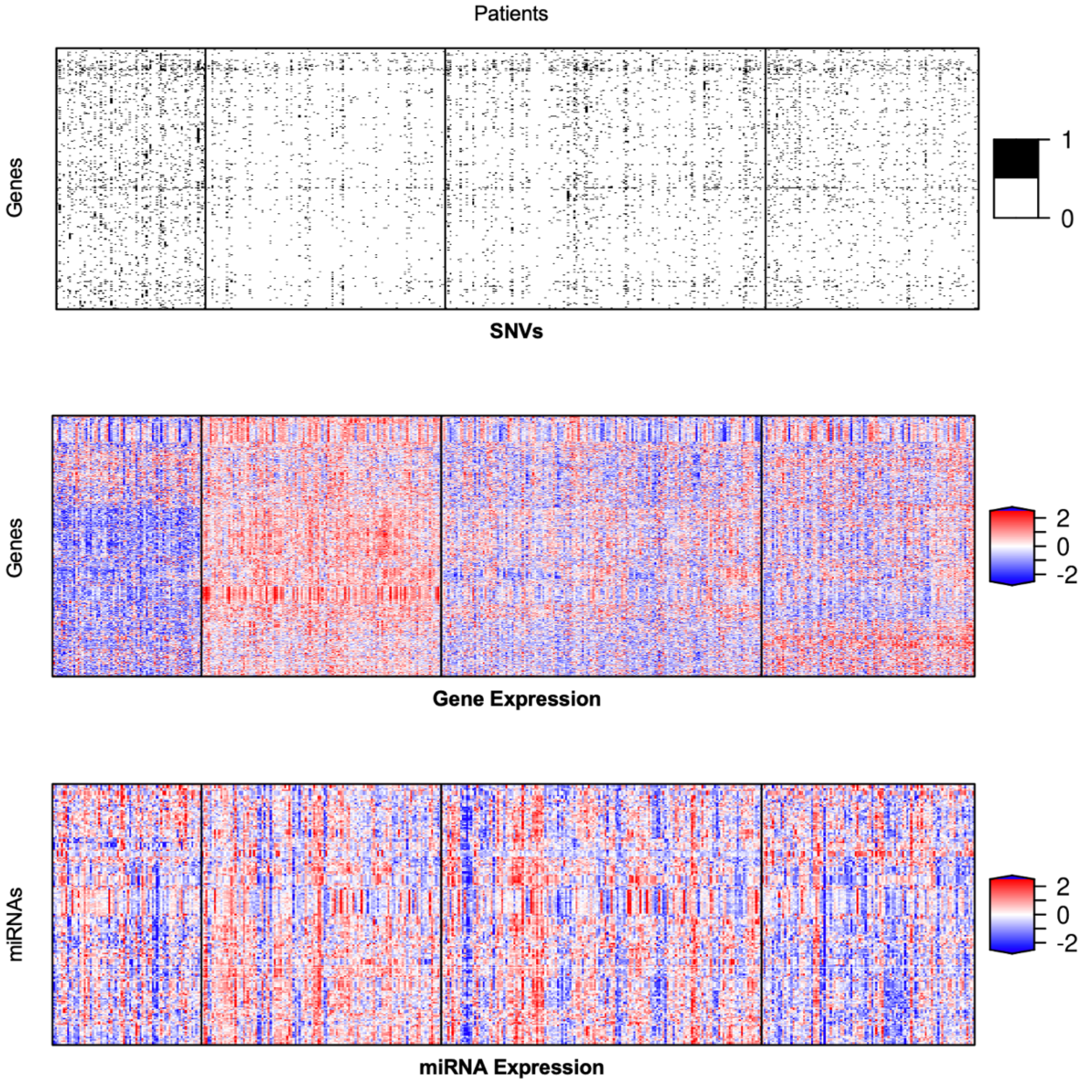


Figure 3.6. Heatmap showing LUAD samples (columns) and genomic features (rows). SNV genes (first panel), gene expression (second panel) and miRNA expression (third panel) for the 3 clusters.

3.3.1.1. Feature selection

Genomic features which drive the integrative clustering, were determined using the posterior probability bigger than 0.5 for SNV, gene expression and miRNA expression. Gene and pathway enrichment were performed for the full list of features which have posterior probability (pp) bigger than 0.5. Besides, the features can be filtered by their posterior probability. For example, the top ten significant features which have pp > 0.9 are CSMD3, KRAS, TP53, SPTA1, ZFHX4, STK11, XIRP2, SMARCA4, PCLO, KEAP1, EGFR of 268 SNV genes; GCLC, CFTR, AOC1, KLHL13, SLC7A2, ZMYND10, ARX, PON1, SOX8, ITGA2B of 1788 expressed genes; hsa-let-7a-1, hsa-let-7a-2, hsa-let-7a-3, hsa-let-7b, hsa-let-7c, hsa-let-7f-1, hsa-let-7f-2, hsa-mir-100, hsa-mir-101-1, hsa-mir-101-2 of 186 expressed miRNAs.

3.3.1.2. Pathway Enrichment

According to pathway enrichment from KEGG database, SNV genes are related with cancer related pathways including ABC transporters, tyrosine kinase inhibitor resistance and glioma; metabolic pathways such as protein digestion and central carbon metabolism; and extracellular interaction pathways such as ECM-receptor interaction, focal adhesion and cell adhesion. Gene expression are highly related with immune pathways such as cytokine-cytokine receptor interaction, IL-17 signaling pathway, chemokine signaling pathway and complement and coagulation cascade; metabolic pathways such as arachidonic acid metabolism, protein digestion and absorption, linoleic acid metabolism, lipid and nitrogen metabolism; extracellular interaction pathways such as ECM-receptor signaling and cell adhesion; and PI3K-Akt signaling pathway which is a well-known cancer related pathway, while miRNA expression is highly related with miRNAs in cancer and chemical carcinogenesis receptor activation (Figure 3.7).

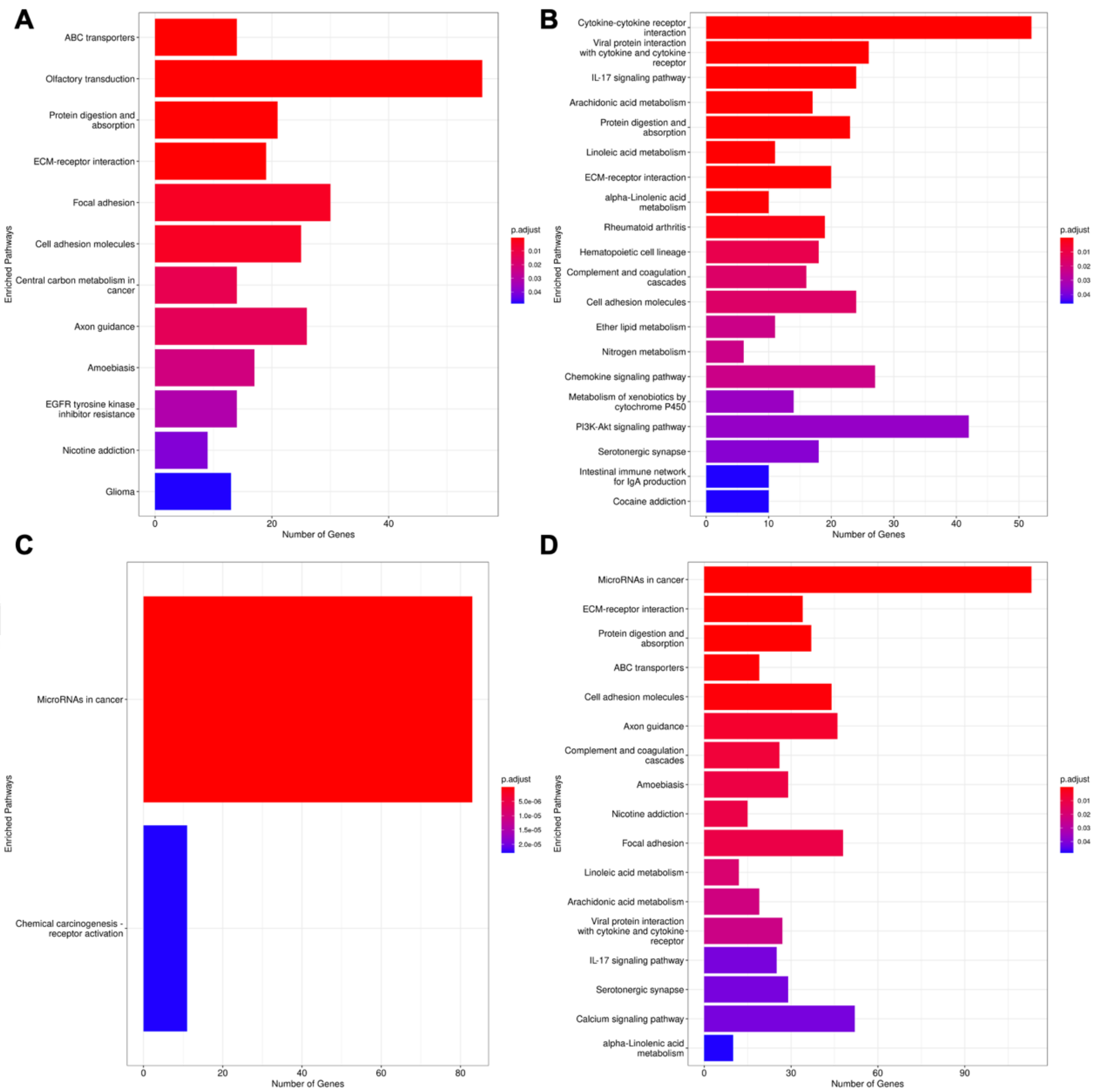


Figure 3.7. Pathways of SNVs, gene expression and miRNA expression used for integrative clustering of LUAD tumor samples. (A) Pathways of genes with SNVs, (B) Pathways of gene expression, (C) Pathways of miRNA expression, (D) Pathways of all features used for integrative clustering by iClusterBayes.

3.3.1.3. Comparison of LUAD iClusters for clinical variables and sample biology

In order to determine the relationship between iClusters and clinical variables and biology of tumor samples, we performed several statistical analyses. Firstly, we performed Pearson's $\chi 2$-test of independence to determine the relationship between iClusters and primary diagnosis. The significant results of the test show that there is a significant association between iClusters and primary diagnosis. According to bar graph, all iClusters mostly consist of adenocarcinoma subtype and iCluster1 has highest percentage of adenocarcinoma samples with absence of acinar cell carcinoma samples. On the other hand, iCluster2 has lowest percentage of adenocarcinoma samples with higher percentage of mucinous and mixed subtypes (Figure 3.8 A). There are not significant associations between iClusters and tumor stage or new tumor event type although iCluster2 has highest percentage of stage I tumor samples and lowest percentage of distant metastasis. iCluster1 and iCluster4 have lower percent of stage I and higher percent of stage III tumor samples while they have higher percent of distant metastasis. There is a significant association between iClusters and treatment outcome of first drug course. iCluster2 patients gave a complete response with highest ratio while iCluster1 patients gave lowest amount of complete response and highest amount of progressive disease response (Figure 3.8).

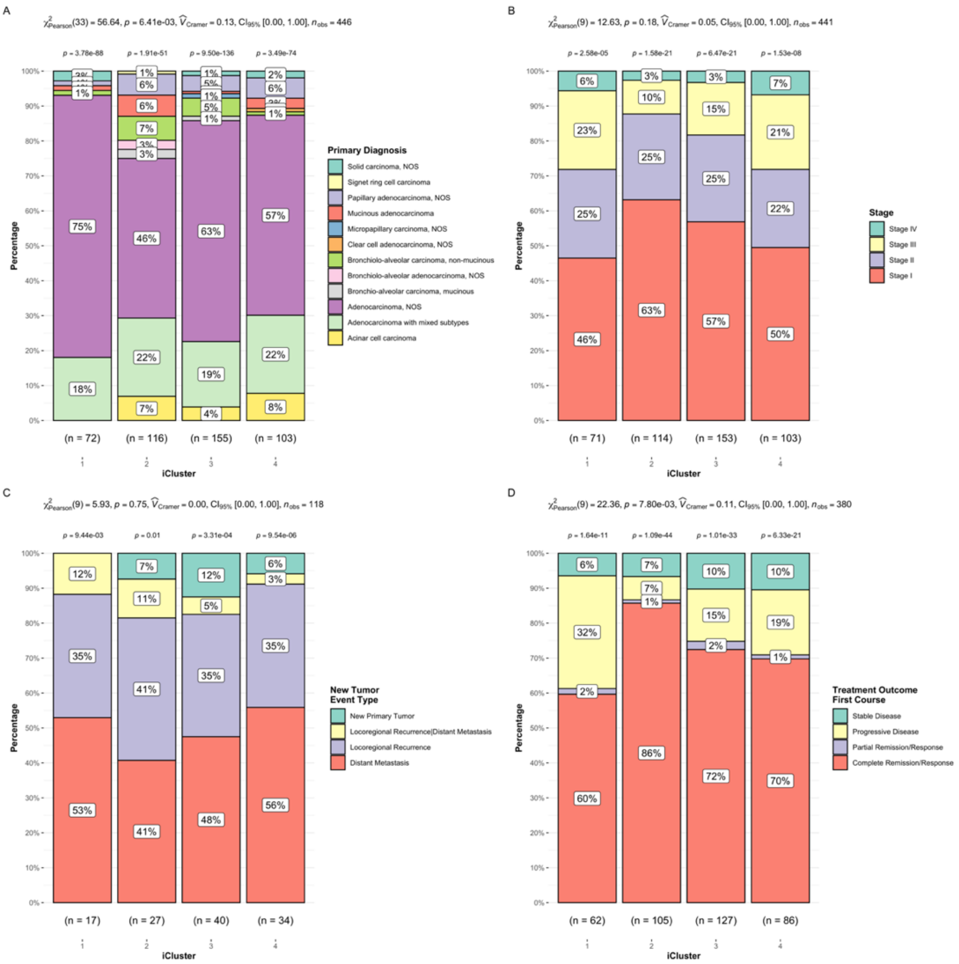


Figure 3.8. Relationships between iClusters and clinical variables of LUAD patients. (A) Bar chart of iClusters vs primary diagnosis (A), tumor stage (B), new tumor event type (C), treatment outcome of first course (D) and results of Pearson's $\chi 2$-test for each.

In order to compare the clinical variables of patients from different iClusters we analyzed different survival and interval probabilities. When we analyzed survival probability of 4 LUAD patient clusters, there was significant difference between the iClusters for overall survival (OS), progression-free interval (PFI) and disease-free interval (DSI). According to plots, iCluster1 patients have worse prognosis while iCluster2 patients have better prognosis (Figure 3.9).

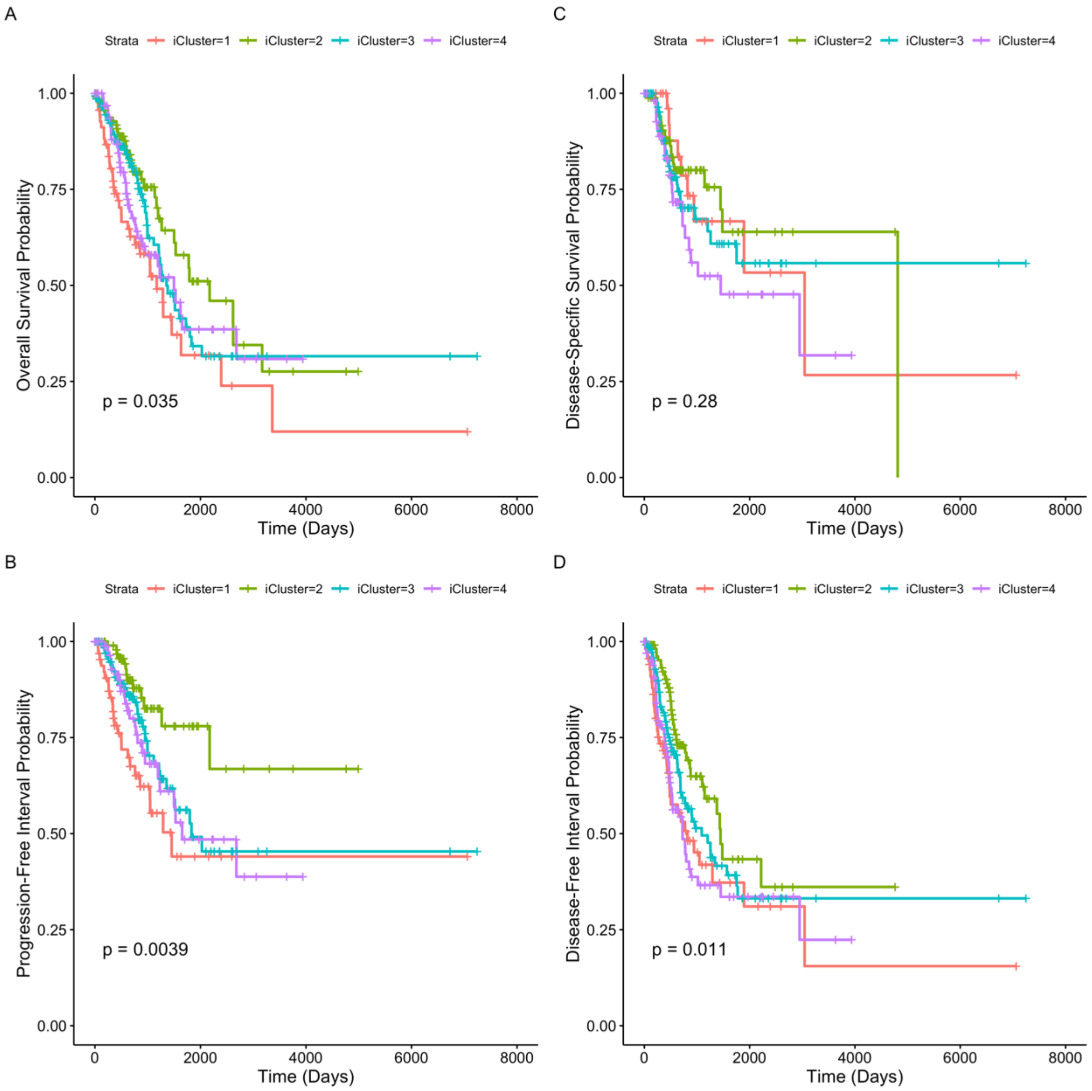


Figure 3.9. Survival and interval analysis results of LUAD iClusters.

We compared iClusters with TCGA subtypes which are estimated by TCGA research network and published in previous articles. At the heatmap below, it is seen that there are associations between expression subtypes, integrative subtypes, immune subtypes and iClusters (Figure 3.10).

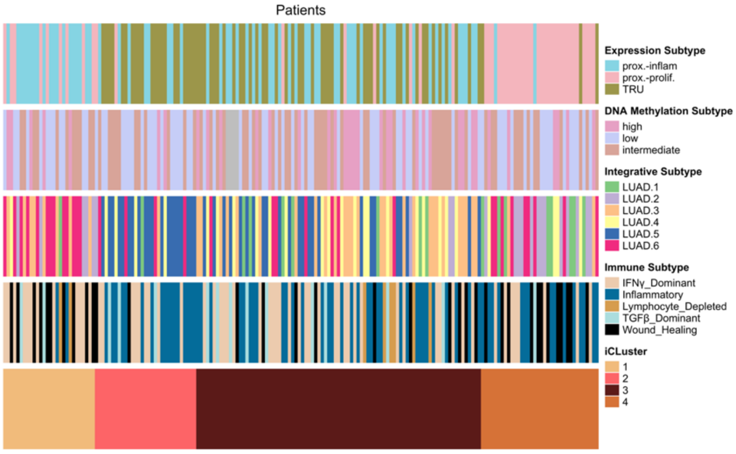


Figure 3.10. Heatmap comparison of LUAD iClusters with TCGA subtypes curated from publications in literature.

When we performed statistical analyses to determine the associations, we saw that iClusters are significantly associated with expression subtypes, integrative subtypes and immune subtypes. The associations between expression subtypes, integrative subtypes and iClusters can be explainable dependent on the similar methodology used by researchers. iClusters are highly dependent on gene expression and their association with expression subtype is highly significant (Figure 3.11). We want to give our attention to immune subtypes because they may be related with response of the patients to immunotherapeutics. iCluster2 has the highest percent of inflammatory subtype while iCluster1 has the lowest percent of inflammatory subtype and higher amount of wound healing subtype with iCluster4.

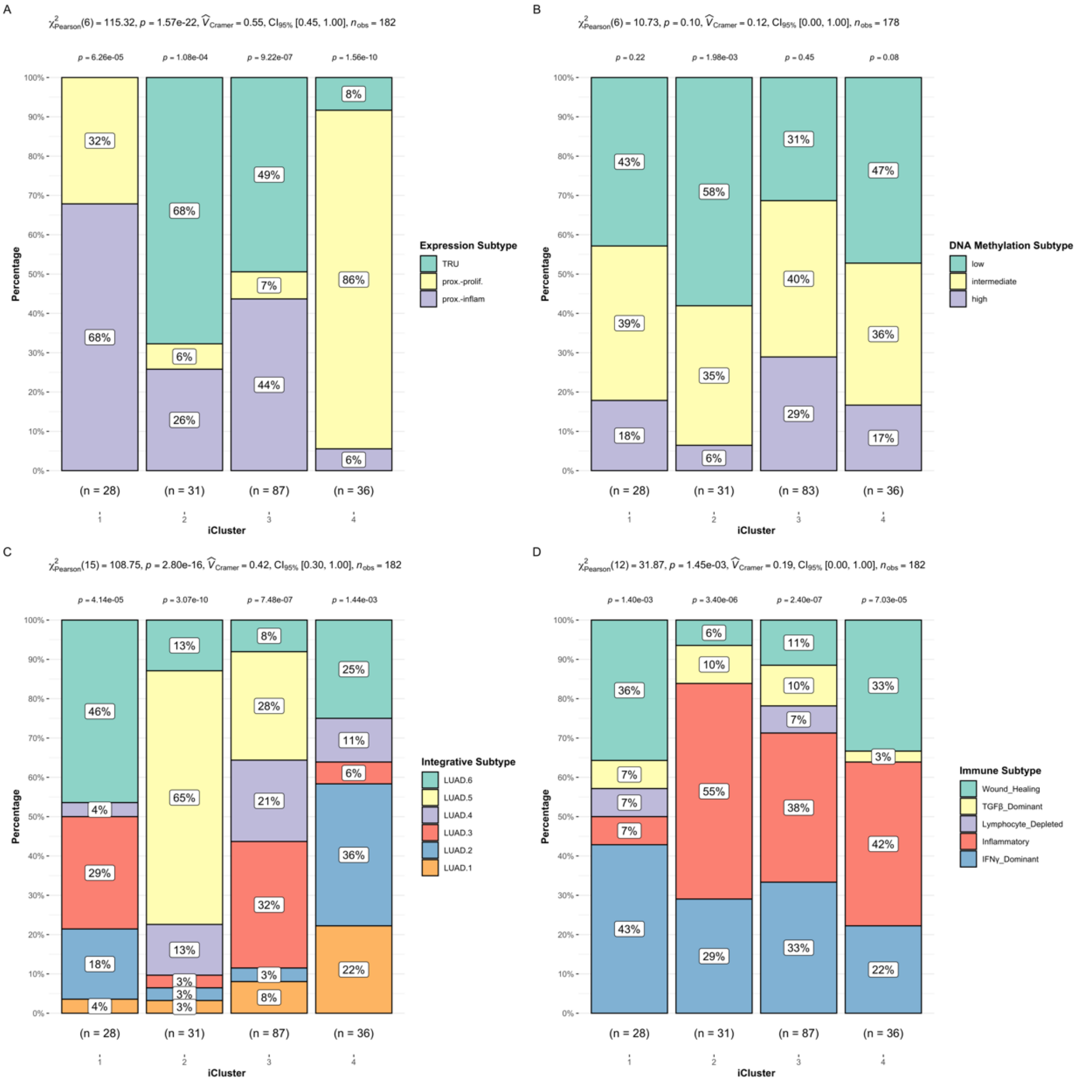


Figure 3.11. Relationships between LUAD iClusters and TCGA subtypes. (A) Bar chart of iClusters vs expression subtype (A), DNA methylation subtype (B), Integrative subtype (C), immune subtype (D) and results of Pearson's $\chi 2$-test for each.

We wanted to test the relationship between immune subtypes and overall survival of patients because the higher inflammatory amount of iCluster2 may be related to better prognosis of patients in iCluster2. According to survival analysis, inflammatory subtype has higher overall survival probability than the other subtypes and wound healing subtype has the lowest survival probability followed by lymphocyte depleted subtype (Figure 3.12).

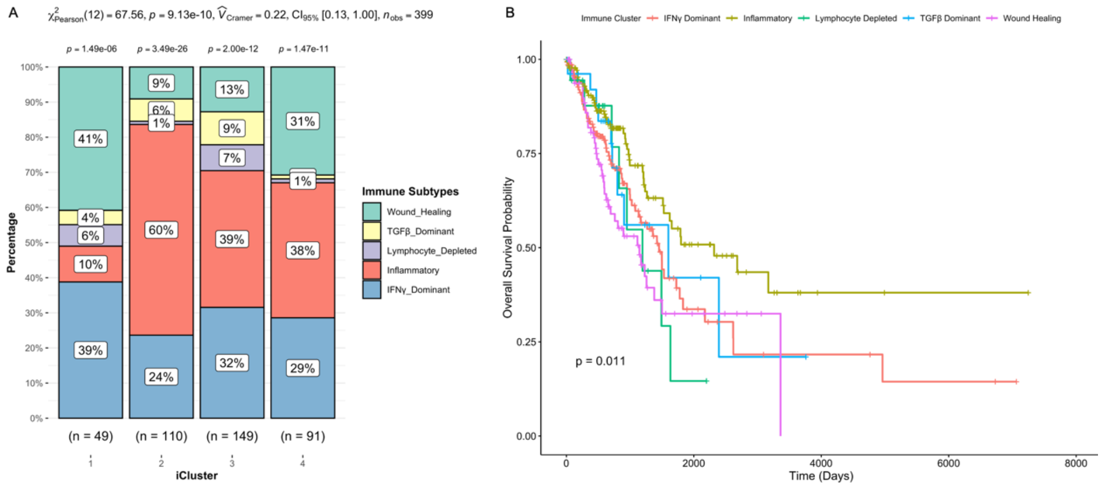


Figure 3.12. Relationships between LUAD iClusters and immune subtypes. (A) Bar chart of iClusters vs immune subtypes and results of Pearson's $\chi 2$-test, (B) Overall survival analysis of immune subtypes.

In order to determine immune related nature of iClusters, we analyzed the purity ratio of tumor samples, fraction of tumor infiltrated leukocytes, fraction of CD4+ and CD8+ tumor infiltrated T cells in tumor samples of iClusters. There are significant differences between iClusters in terms of purity, tumor infiltrated leukocytes and fraction of CD4+ tumor infiltrated T cells. iCluster2 has lower purity ratio with higher fraction of tumor infiltrated leukocytes and fraction of CD4+ tumor infiltrated T cells (Figure 3.13).

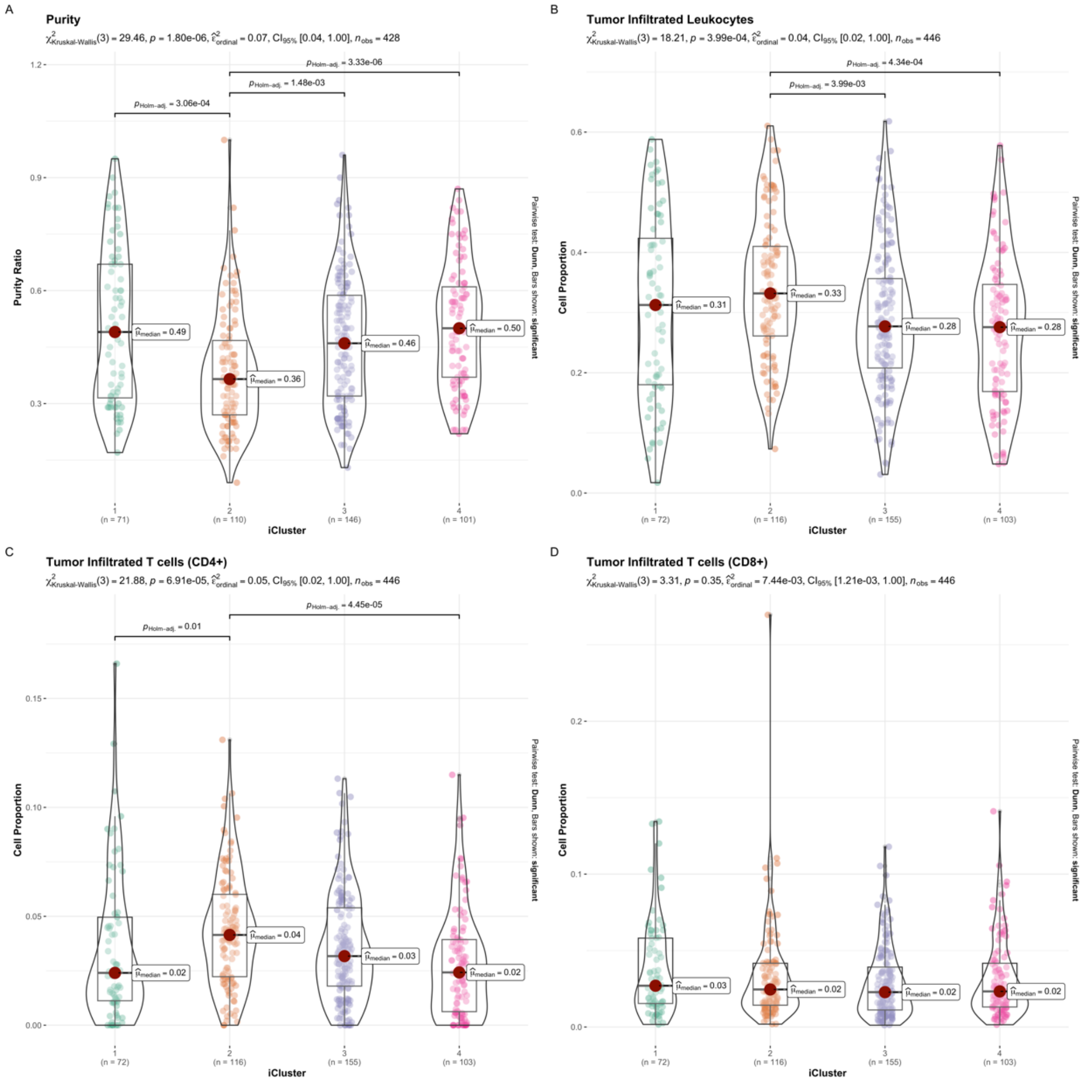


Figure 3.13. Comparison of LUAD iClusters for tumor infiltrated immune cells. (A) Purity ratio of tumor samples, (B) Fraction of tumor infiltrated leukocytes, (C) Proportion of CD4+ tumor infiltrated T cells, (D) Proportion of CD8+ tumor infiltrated T cells.

We also determined the differences of tumor mutation burden/load (TMB) and homologous DNA repair deficiency (HRD) level among iClusters. There are significant differences between iClusters for TMB and HRD. iCluster1 has the highest amount of TMB and HRD score followed by iCluster4 while iCluster2 has the opposite situation significantly (Figure 3.14).

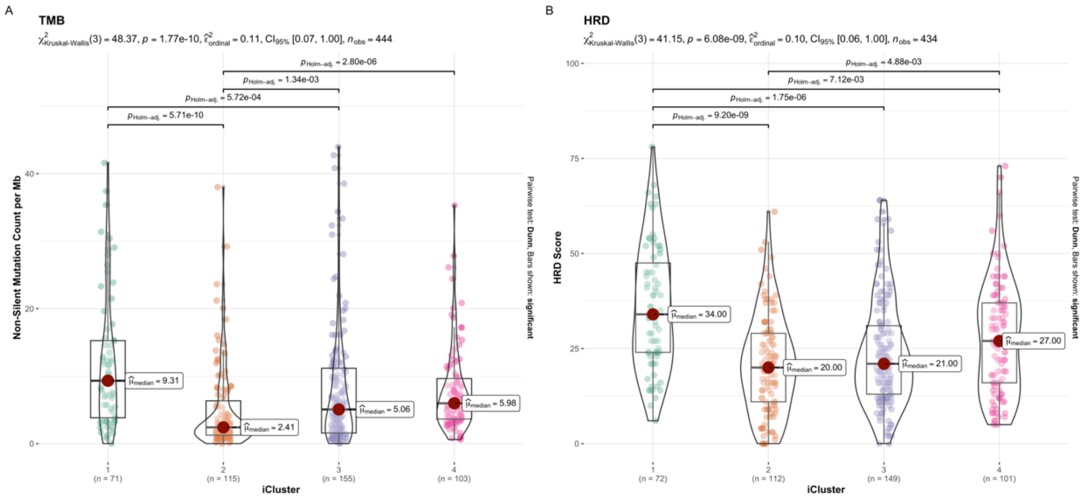


Figure 3.14. Comparison of LUAD iClusters for Tumor Mutation Burden (A) and Homologous DNA Repair Deficiency (B).

In order to see the correlation between TMB/HRD and tumor infiltrated immune cell fractions independent of iClusters, we analyzed the correlation matrix between them. Firstly, TMB and HRD are positively correlated as expected. Fraction of tumor infiltrated leukocytes (TIL) is not significantly correlated with HRD and TMB, however both HRD and TMB are negatively correlated with CD4+ tumor infiltrated T cells and positively correlated with CD8+ T cells (Figure 3.15).

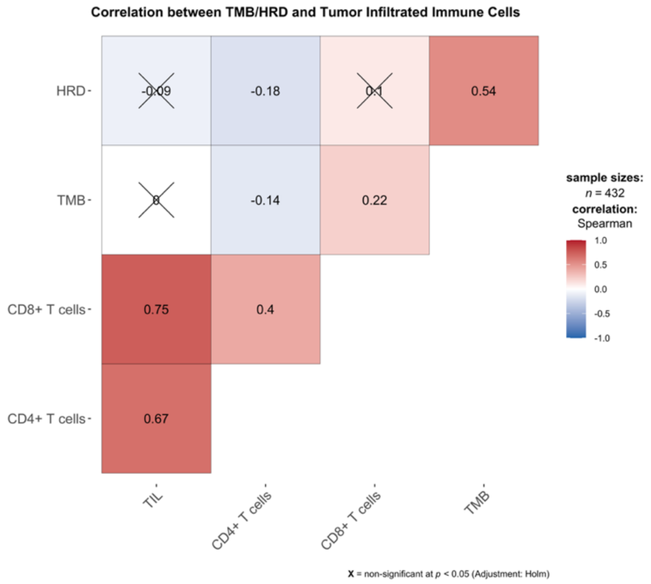


Figure 3.15. Correlation matrix between TMB/HRD and tumor infiltrated immune cells in LUAD tumor samples. TIL: Tumor Infiltrated Leukocytes, TMB: Tumor Mutation Burden, HRD: Homologous DNA Repair Deficiency.

We aimed to analyze expression differences of tumor associated proteins (TAPs) targetable by antibody therapeutics, among iClusters of LUAD patients. According to heatmap below, there are big differences in expression of TAPs between tumor samples and patients may have different expression patterns of TAPs such as higher expression of all genes, lower expression of all genes or controversial expression of genes for example some patients have higher expression of EGFR while have lower expression of KDR (VEGFR-2) and VEGFA. The expression pattern of TAPs appears to correlate with iClusters but not with risk groups (Figure 3.16).

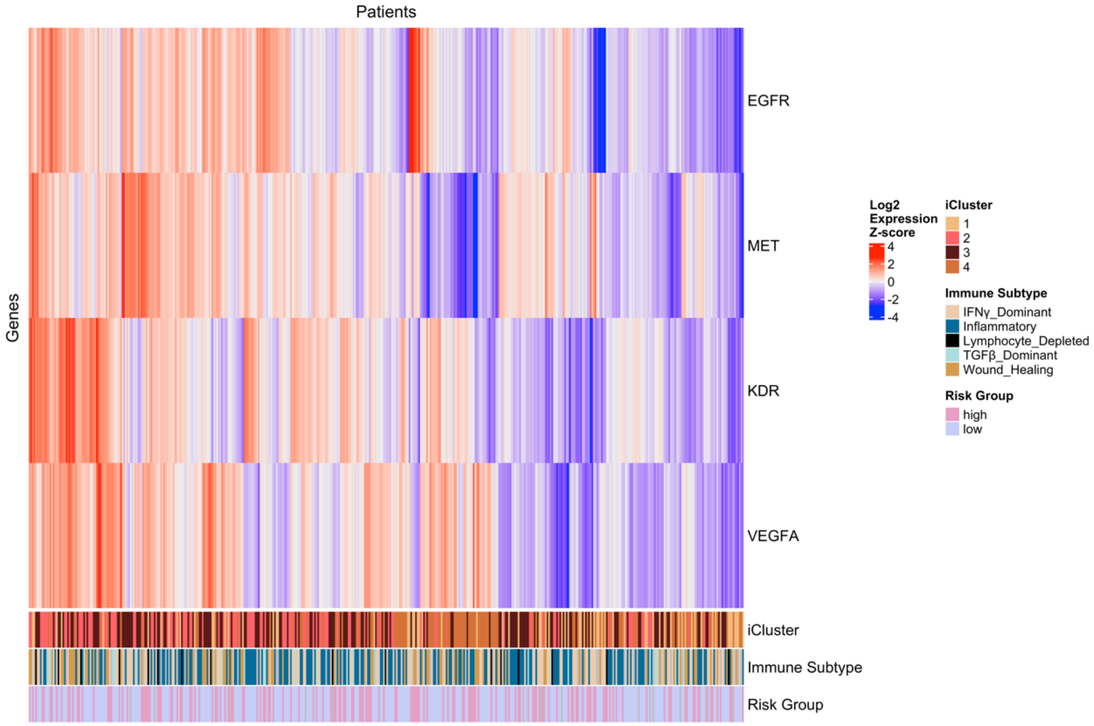


Figure 3.16. Expression heatmap of tumor associated proteins targetable by antibody therapeutics in LUAD tumor samples.

We also analyzed the expression differences of immune associated proteins (IAPs) such as CTLA4, LAG3, PD-1 (PDCD1) and PD-L1 (CD274) which are targetable by antibody therapeutics. According to expression heatmap, there are big differences in expression of IAPs between tumor samples and patients may have different expression patterns of IAPs. While CTAL4 and LAG3 have close expression pattern, some patients have opposite expression of PD-L1 (CD274) and CTLA4 or LAG3.

The expression pattern of IAPs appears to correlate with iClusters but not with risk groups (Figure 3.17).

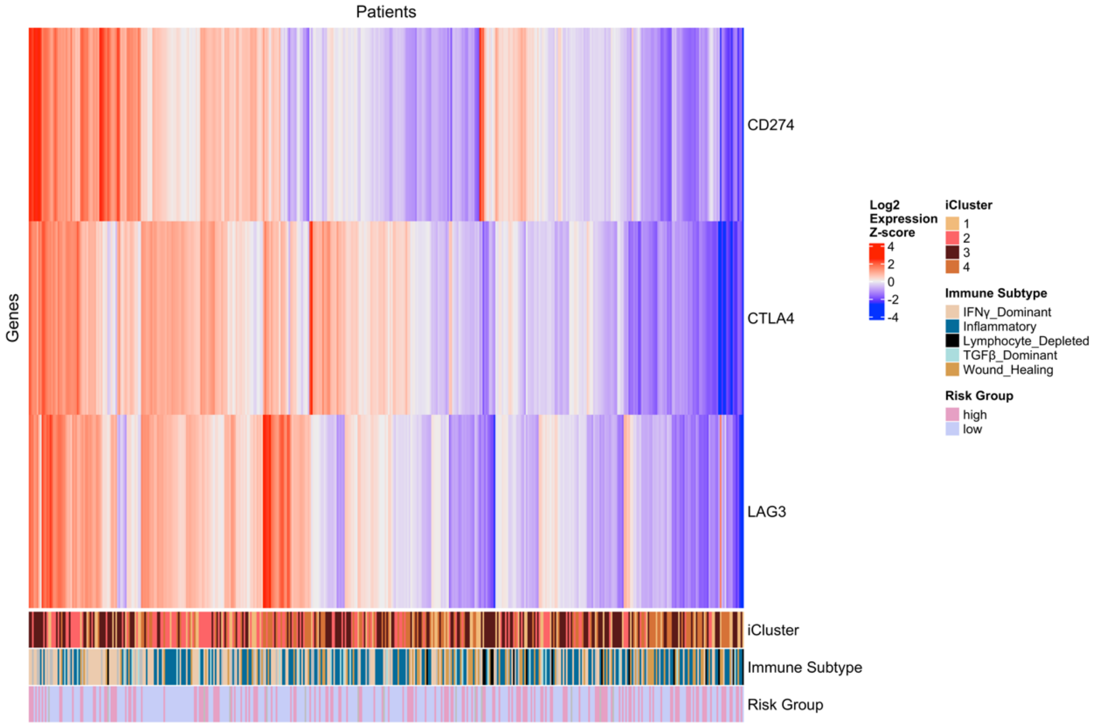


Figure 3.17. Expression heatmap of immune associated proteins targetable by antibody therapeutics in LUAD tumor samples.

In order to identify the relationships between iClusters and expression of TAPs and IAPs, we performed continuous variable comparison across iClusters. Violin plots are showing significant differences between iClusters for CD274, CTLA4, EGFR, KDR, MET and VEGFA genes. iCluster1 and iCluster4 have significantly lower expression of IAPs and most of the TAPs while iCluster2 has higher expression of all TAPs and IAPs (Figure 3.18).

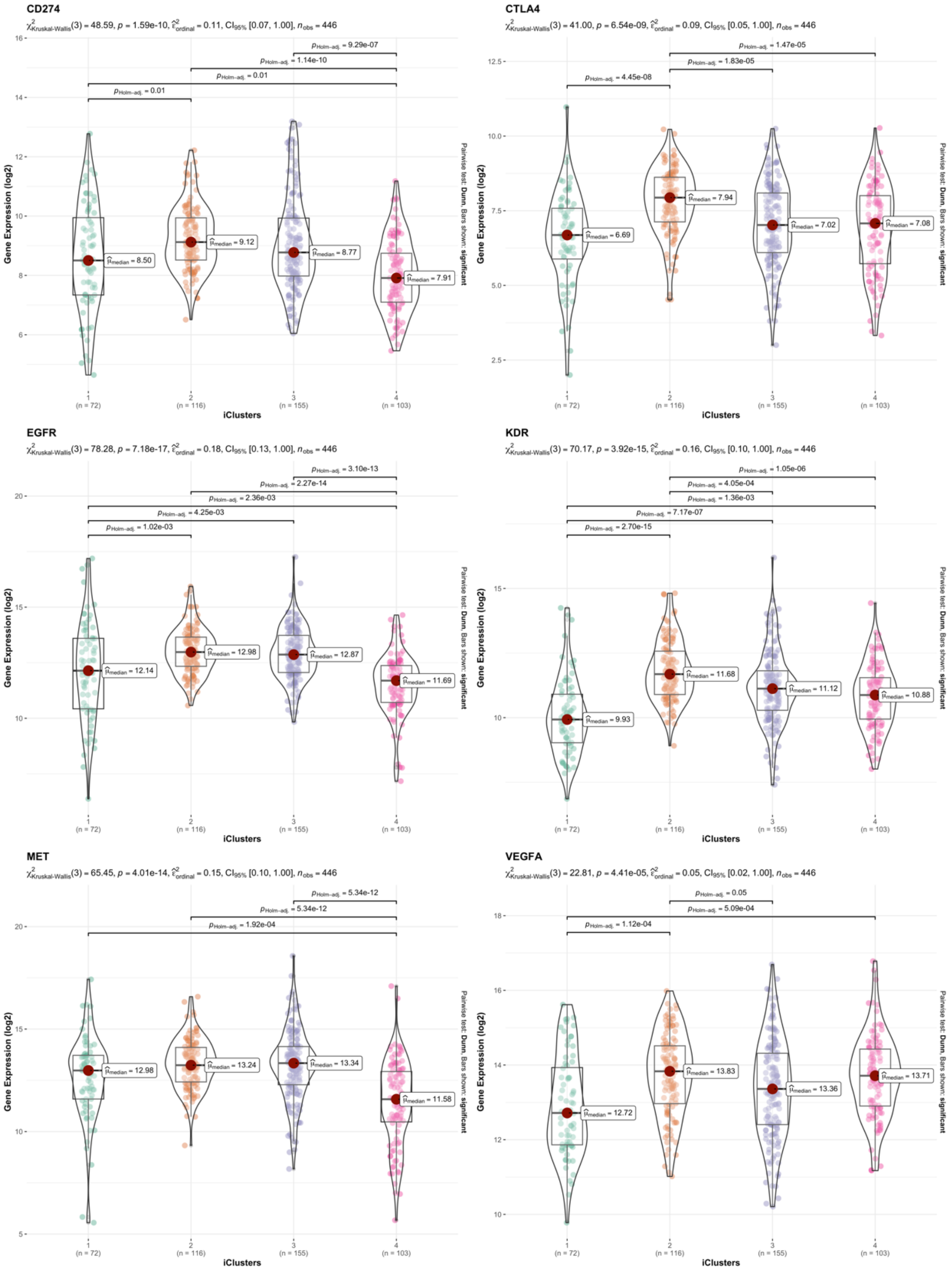


Figure 3.18. Comparison of LUAD iClusters for expression of targetable proteins by antibody therapeutics.

In heatmaps of TAP and IAP expression, risk clusters do not appear to be associated with expression patterns. We analyzed the relationship between iClusters, expression of targetable genes and risk clusters. Pearson's $\chi 2$-test of independence showed that there is a significant association between iClusters and risk clusters (p=5.5E-5). According to bar graph below, iCluster2 has higher percent of low-risk patients while iCluster1 has the highest percent of high-risk patients (Figure 3.19).

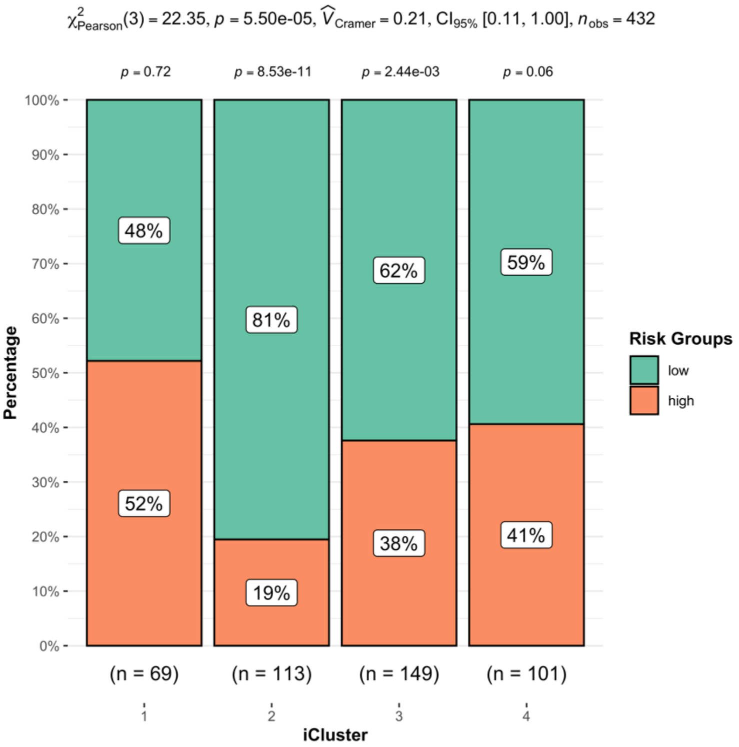


Figure 3.19. Relationships between iClusters and risk clusters of LUAD patients. Bar chart of iClusters vs risk clusters and results of Pearson's $\chi 2$-test.

When we checked the expression levels of immunotherapeutic targetable genes among risk groups, there is not significant differences between risk groups for all genes except CTLA4. Expression of CTLA4 is significantly higher in low-risk group (Figure 3.20).

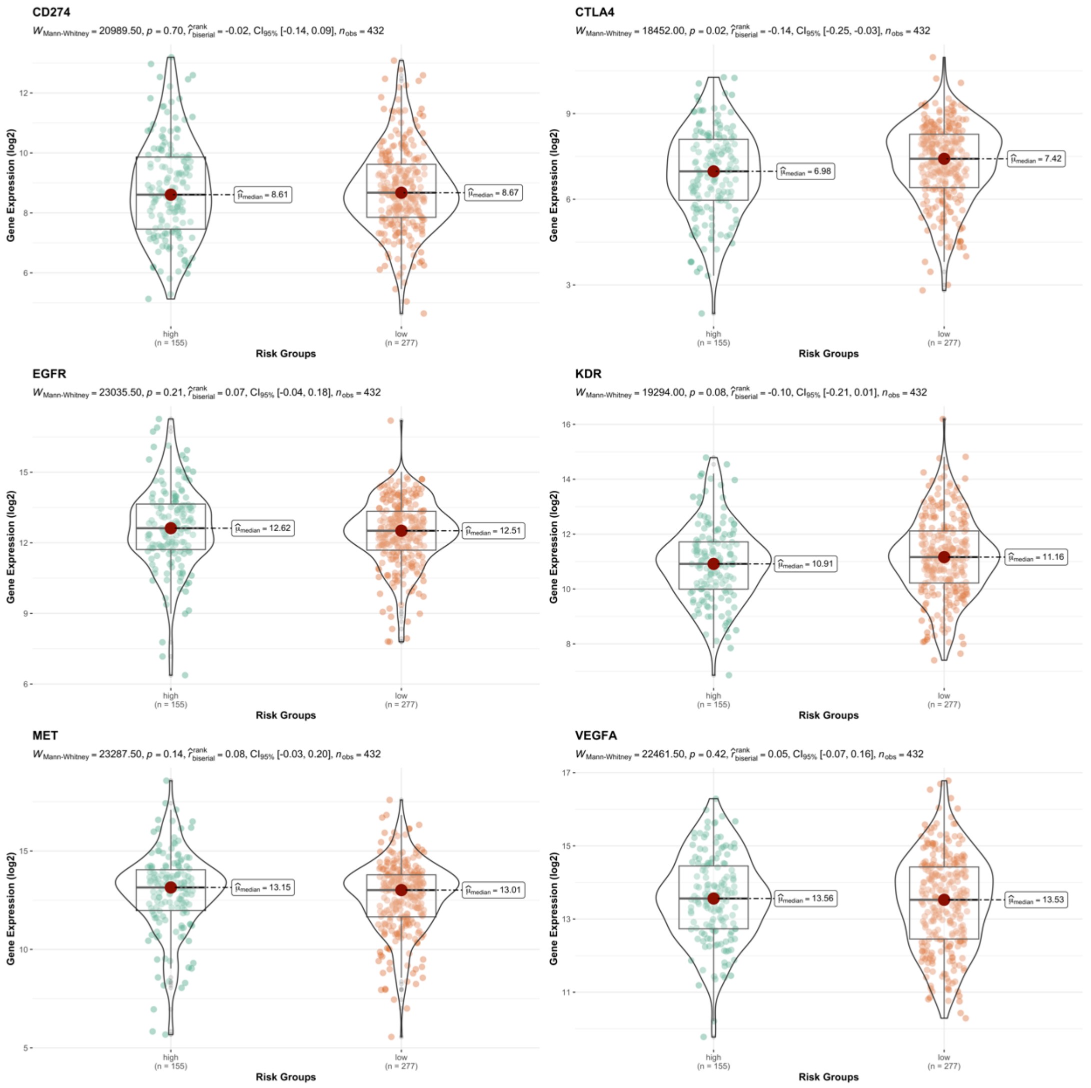


Figure 3.20. Comparison of LUAD risk clusters for expression of targetable proteins by antibody therapeutics.

3.3.2. Integrative clustering of LUSC patients

We analyzed 359 Lung squamous cell carcinoma (LUSC) samples using SNV, CNV, gene expression, miRNA expression and methylation data. We used genes which have >1% somatic mutation rate, non-redundant copy number regions, log2-transformed normalized top %10 most-variable gene expression data, all miRNA expression after low count filter, and top 1% most-variable DNA methylation after NA removing. From the BIC and deviance ratio plots, we saw that k=2 (k+1=3 clusters) is the most optimal solution (Figure 3.21).

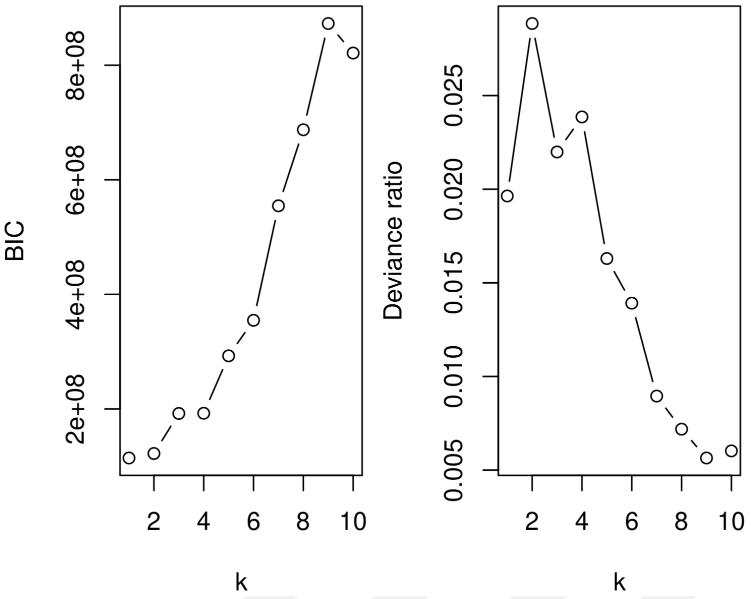


Figure 3.21. Bayesian Information Criteria (BIC) and the deviance ratio vs k for LUSC dataset.

iClusterBayes analysis showed that the number of genes containing SNVs is 707, the number of CNV genes are 7, while the number of expressed genes is 1821 and the number of expressed miRNA genes is 163, which have posterior probabilities >0.5. However, there are no methylation site which have posterior probabilities >0.5, therefore methylation data is not informative for clustering for this dataset. According to the plot of posterior probabilities, gene expression is highly dominant to drive integrative clustering followed by SNVs and miRNAs while the methylation data is not informative for clustering (Figure 3.22).

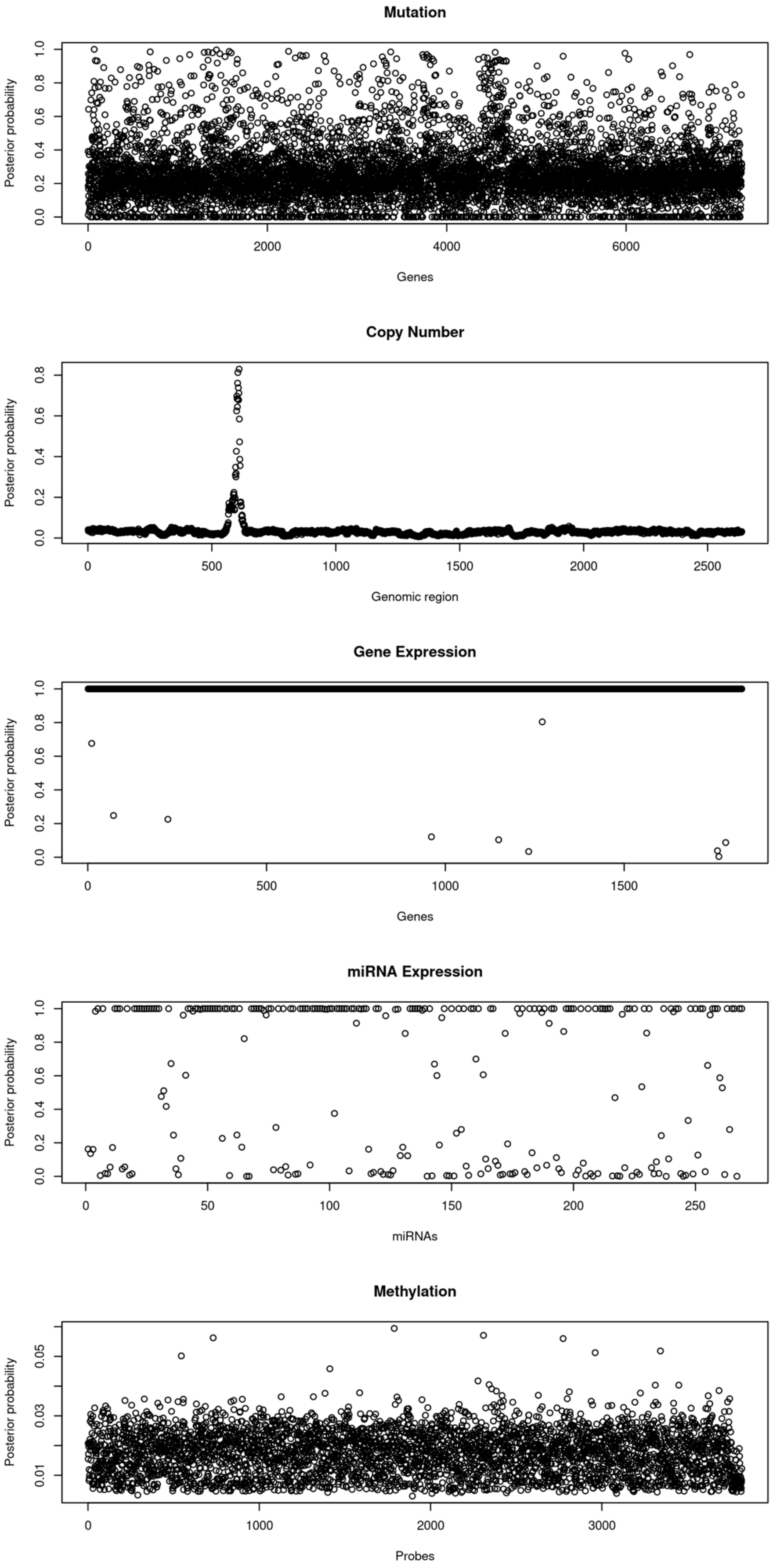


Figure 3.22. Posterior probabilities of genomic features that drive the integrative clustering.

According to the heatmap, there is significant differences between clusters. For CNVs, iCluster2 has higher copy number of a segment on chromosome 3. For gene expression, iCluster3 has higher expression of same set of genes. SNVs and miRNA expression patterns have differences between clusters (Figure 3.23).

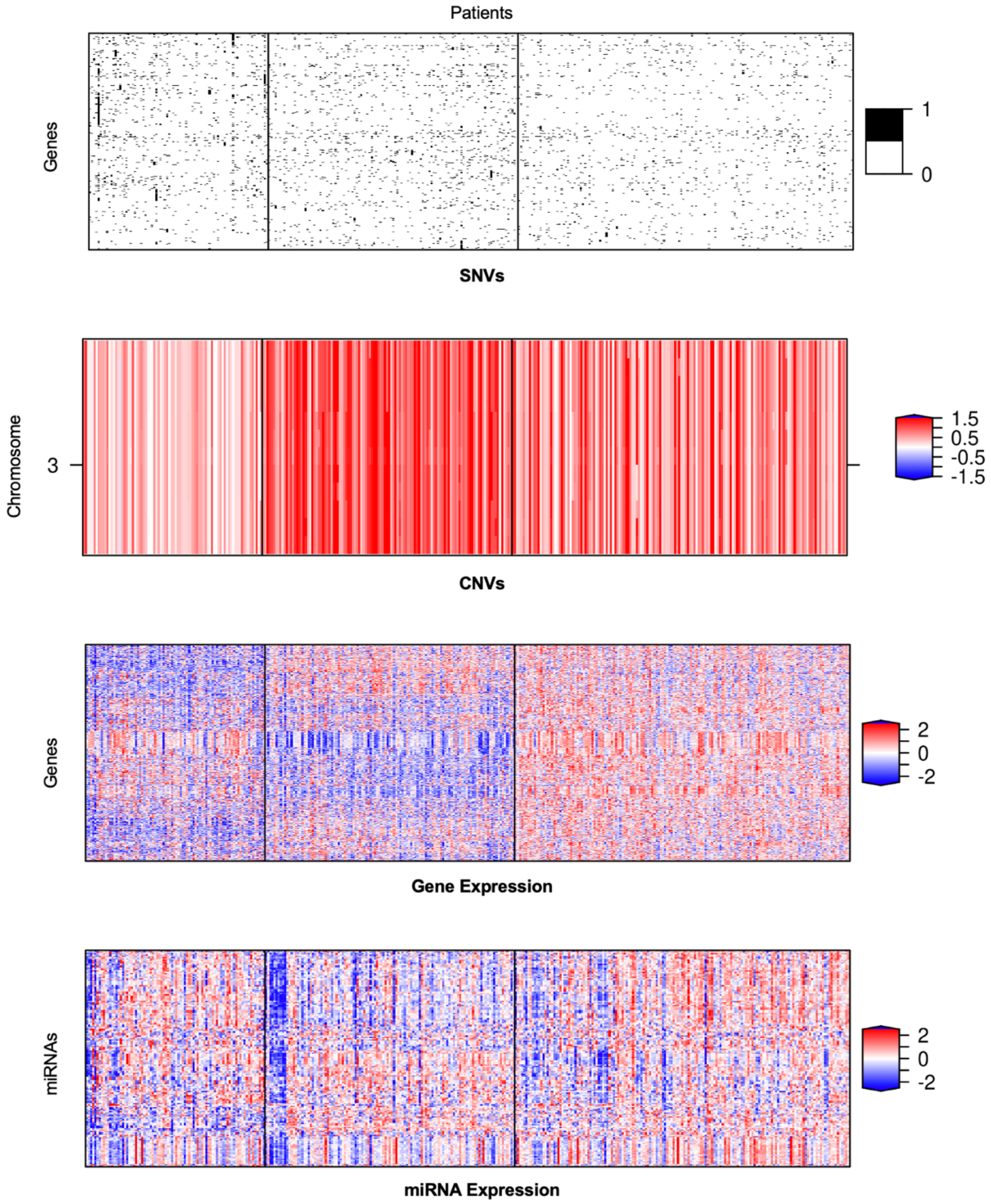


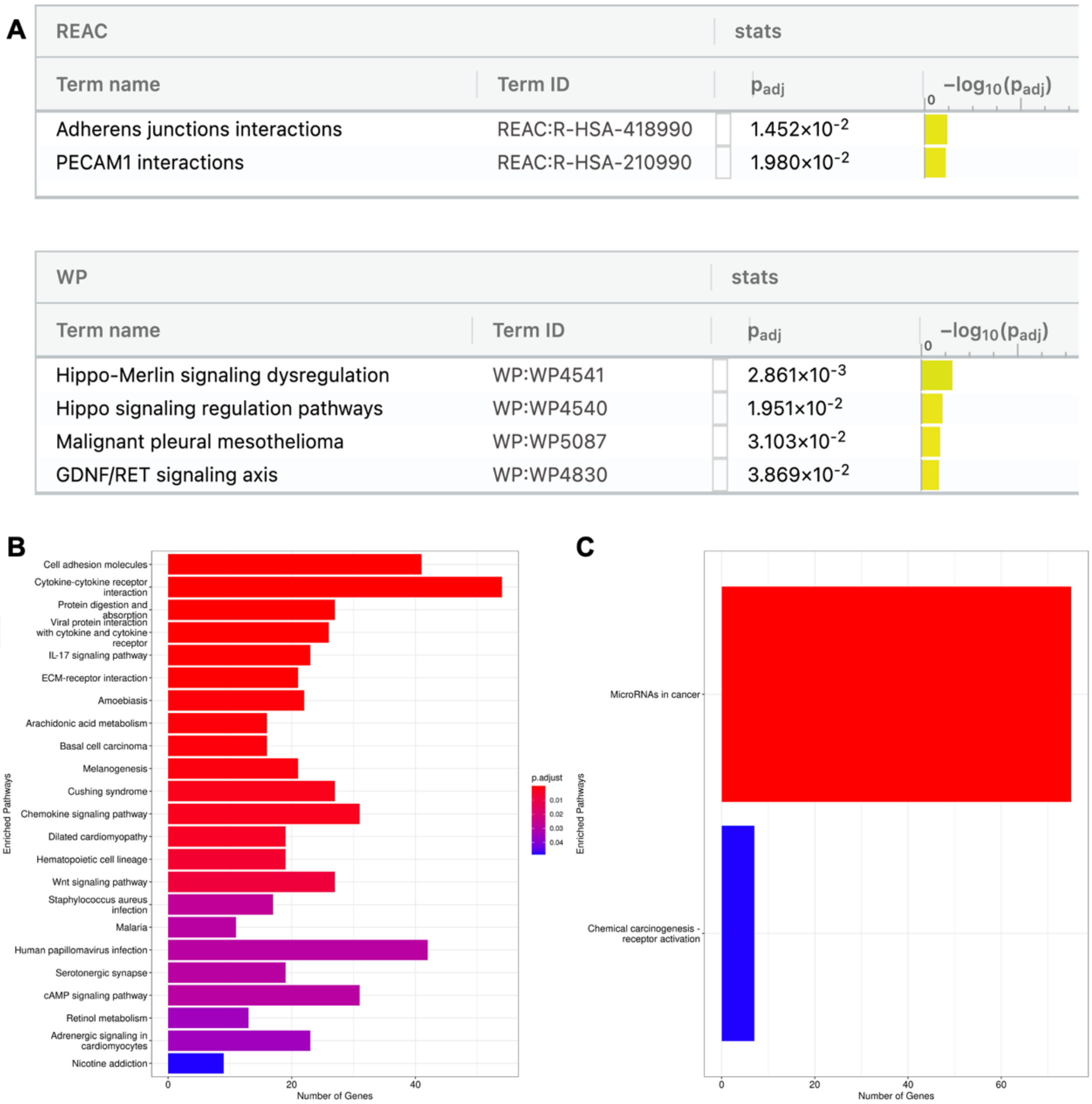
Figure 3.23. Heatmap showing LUSC samples (columns) and genomic features (rows). SNV genes (first panel), CNV regions (second panel), gene expression (third panel) and methylation data (fourth panel) for the 3 clusters.

3.3.2.1. Feature selection

Genomic features which drive the integrative clustering of LUSC patients, were determined using the posterior probability bigger than 0.5 for SNVs, CNVs, gene expression and miRNA expression. Gene and pathway enrichment were performed for the full list of features which have posterior probability (pp) bigger than 0.5. Besides, the features can be filtered by their posterior probability. For example, the top ten significant features which have pp > 0.9 are NFE2L2, ENAM, MAP3K15, PTCHD2, FOXD4L4, COL5A1, HECW2, OR13F1, CDH2, and FLG of 65 SNV genes; GCLC, CFTR, AOC1, WNT16, ICA1, SLC7A2, PDK4, ZMYND10, HOXA11, and MEOX1 of 1819 expressed genes; hsa-let-7c, hsa-let-7e, hsa-mir-100, hsa-mir-101-1, hsa-mir-101-2, hsa-mir-106a, hsa-mir-10a, hsa-mir-10b, and hsa-mir-1180 of 147 expressed miRNAs.

3.3.2.2. Pathway Enrichment

According to pathway enrichment analysis, SNV genes are not enriched from KEGG database significantly, however, they play role in adherens junction interactions and PECAM1 interactions according to Reactome database and in hippo signaling regulation, GDNF/RET signaling and malignant pleural mesothelioma pathways according to WikiPathways database. RNA expression genes are highly related with extracellular interaction pathways such as cell adhesion and ECM-receptor signaling; immune pathways such as cytokine-cytokine receptor interaction, IL-17 signaling, chemokine signaling, infection related pathways and hematopoietic cell lineage; metabolic pathways such as protein digestion and absorption, arachidonic acid metabolism; and cancer related pathways such as basal cell carcinoma and Wnt signaling pathway. Lastly, miRNA genes are highly related with miRNAs in cancer and chemical carcinogenesis (Figure 3.24).



A

REAC		stats	
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$
Adherens junctions interactions	REAC:R-HSA-418990	1.452×10^{-2}	
PECAM1 interactions	REAC:R-HSA-210990	1.980×10^{-2}	

WP		stats	
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$
Hippo-Merlin signaling dysregulation	WP:WP4541	2.861×10^{-3}	
Hippo signaling regulation pathways	WP:WP4540	1.951×10^{-2}	
Malignant pleural mesothelioma	WP:WP5087	3.103×10^{-2}	
GDNF/RET signaling axis	WP:WP4830	3.869×10^{-2}	

Figure 3.24. Pathways of SNVs, gene expression and miRNA expression used for integrative clustering of LUSC tumor samples. (A) Pathways of genes with SNVs, (B) Pathways of gene expression, (C) Pathways of miRNA expression.

3.3.2.3. Comparison of LUSC iClusters for clinical variables and sample biology

We analyzed the relationship between iClusters and clinical variables and biology of tumor samples, using several statistical analyses. According to results of Pearson's $\chi 2$-test of independence, there is not significant association between primary diagnosis, new tumor event type, treatment outcome of first course and iClusters. However, iCluster1 has highest percent of distant metastasis and new primary tumor. There is significant association between iClusters and tumor stages, iCluster3 has highest percent of stage I and lowest percent of stage II and stage III tumor samples while iClust1 and iClust2 have less amount of stage I and higher amount of stage II and stage III tumor samples (Figure 3.25).

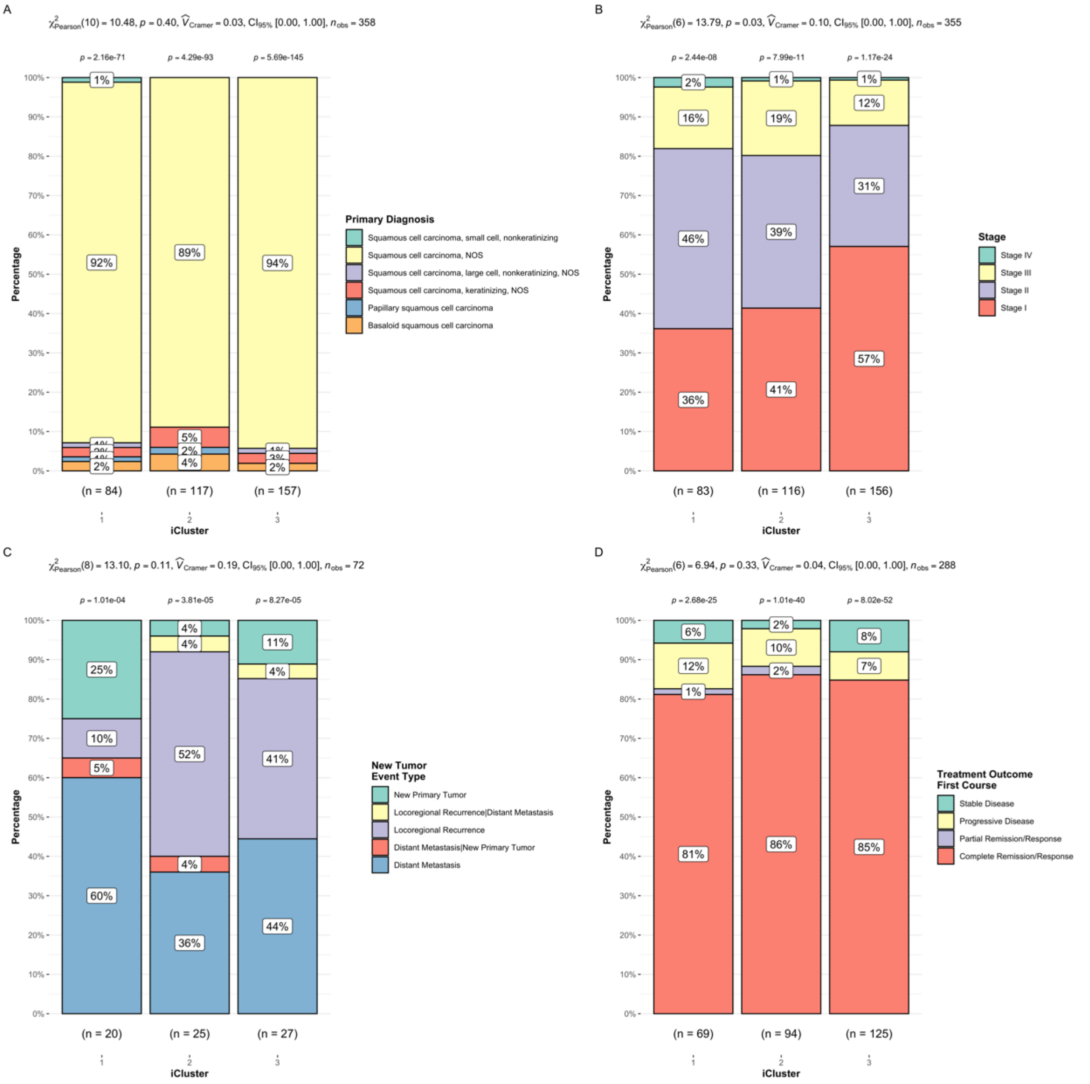


Figure 3.25. Relationships between iClusters and clinical variables of LUSC patients. (A) Bar chart of iClusters vs primary diagnosis (A), tumor stage (B), new tumor event type (C), treatment outcome of first course (D) and results of Pearson's $\chi 2$-test for each.

We analyzed different survival and interval probabilities among iClusters. There is significant difference between the iClusters for only disease-free interval (DSI) but progression-free interval (PFI) is very close to significance level (p=0.059) and has similar pattern with disease-free interval for iCluster1. According to plots, iCluster1 patients have worse prognosis (Figure 3.26).

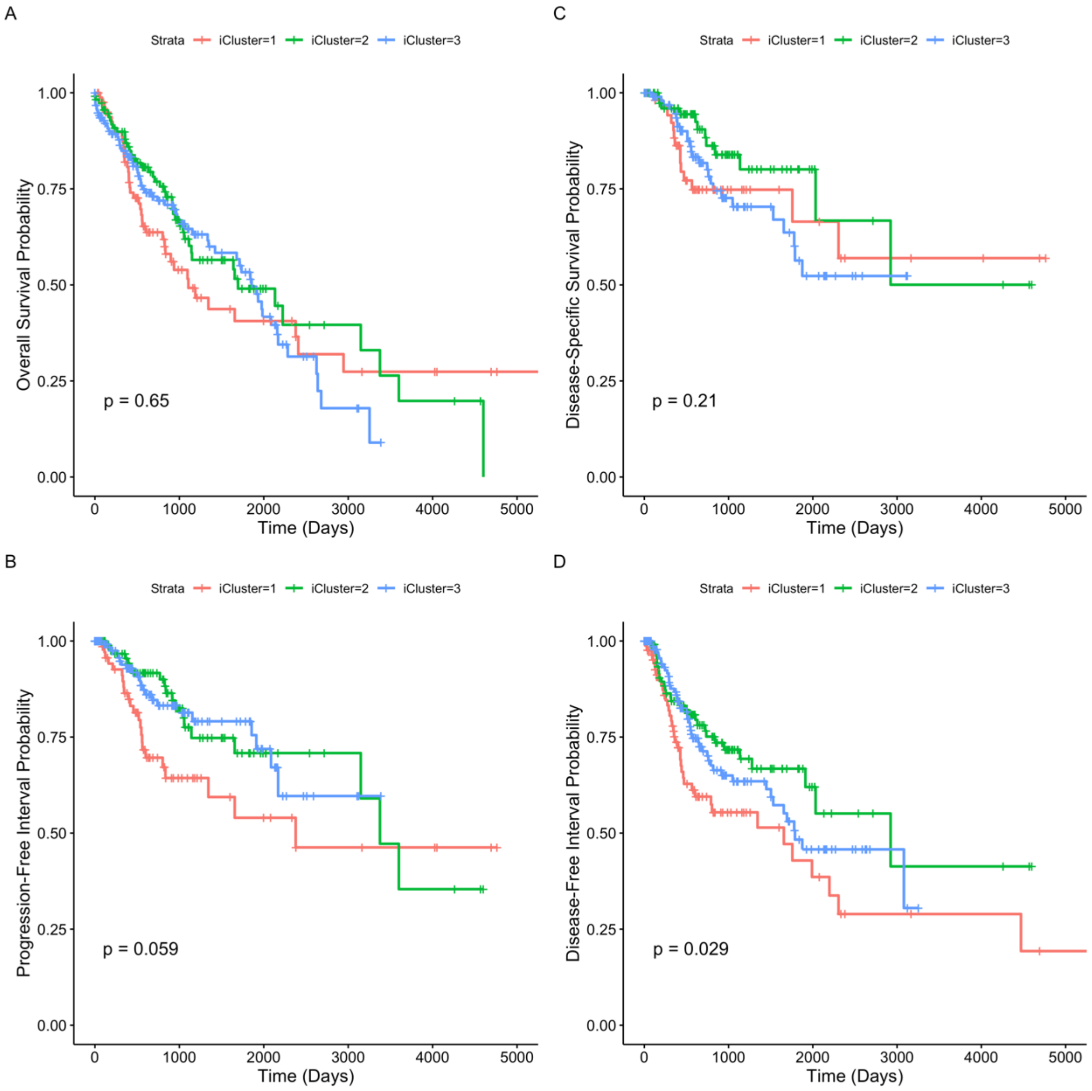


Figure 3.26. Survival and interval analysis results of LUSC iClusters.

We compared iClusters with TCGA subtypes which are estimated by TCGA research network and published in previous articles. At the heatmap below, it is seen that there are associations between expression subtypes, immune subtypes and iClusters (Figure 3.27).

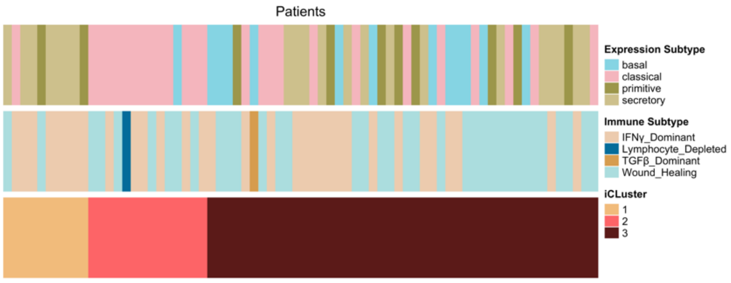


Figure 3.27. Heatmap comparison of LUSC iClusters with TCGA subtypes in literature.

We performed statistical analyses and determined that iClusters are significantly associated with expression subtypes. iCluster1 consists of mostly (70%) secretory, iCluster2 consists of classical (93%) subtype while iCluster3 has almost equal ratios of secretory, primitive, classical and basal expression subtypes. According to relationship with immune subtypes, iCluster1 has the highest percent of IFN-γ dominant subtype while iCluster2 and iCluster3 have mixture of IFN-γ dominant and wound healing subtypes, but the association is not significant (Figure 3.28).

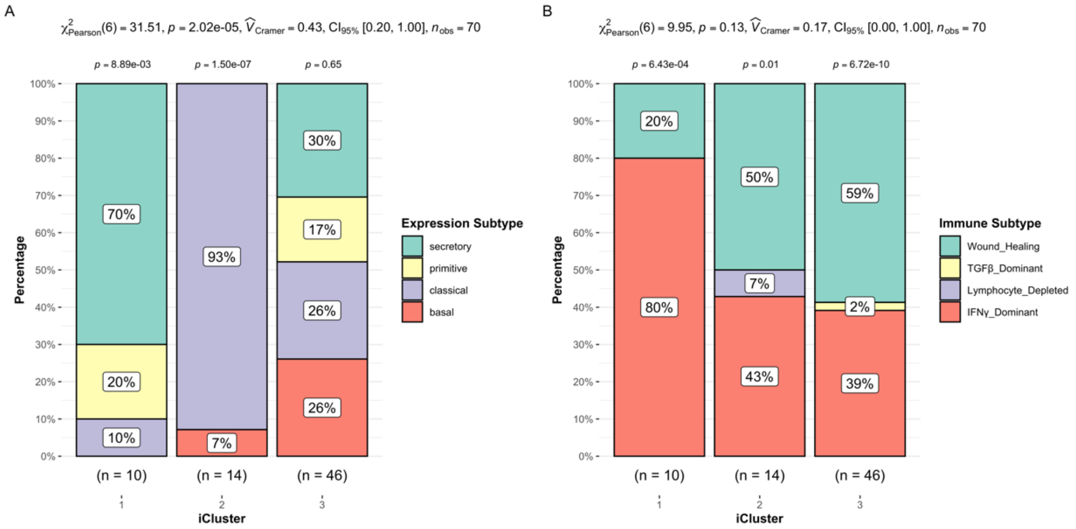


Figure 3.28. Relationships between LUSC iClusters and TCGA subtypes. (A) Bar chart of iClusters vs expression subtype (A), DNA methylation subtype (B), Integrative subtype (C), immune subtype (D) and results of Pearson's $\chi 2$-test for each.

We analyzed the purity ratio of tumor samples, fraction of tumor infiltrated leukocytes, fraction of CD4+ and CD8+ tumor infiltrated T cells in tumor samples of iClusters in order to determine immune related nature of iClusters. There are significant differences between iClusters for all features above. iCluster2 has higher purity ratio with lower fraction of tumor infiltrated leukocytes, lower fraction of CD4+ and CD8+ tumor infiltrated T cells, significantly (Figure 3.29).

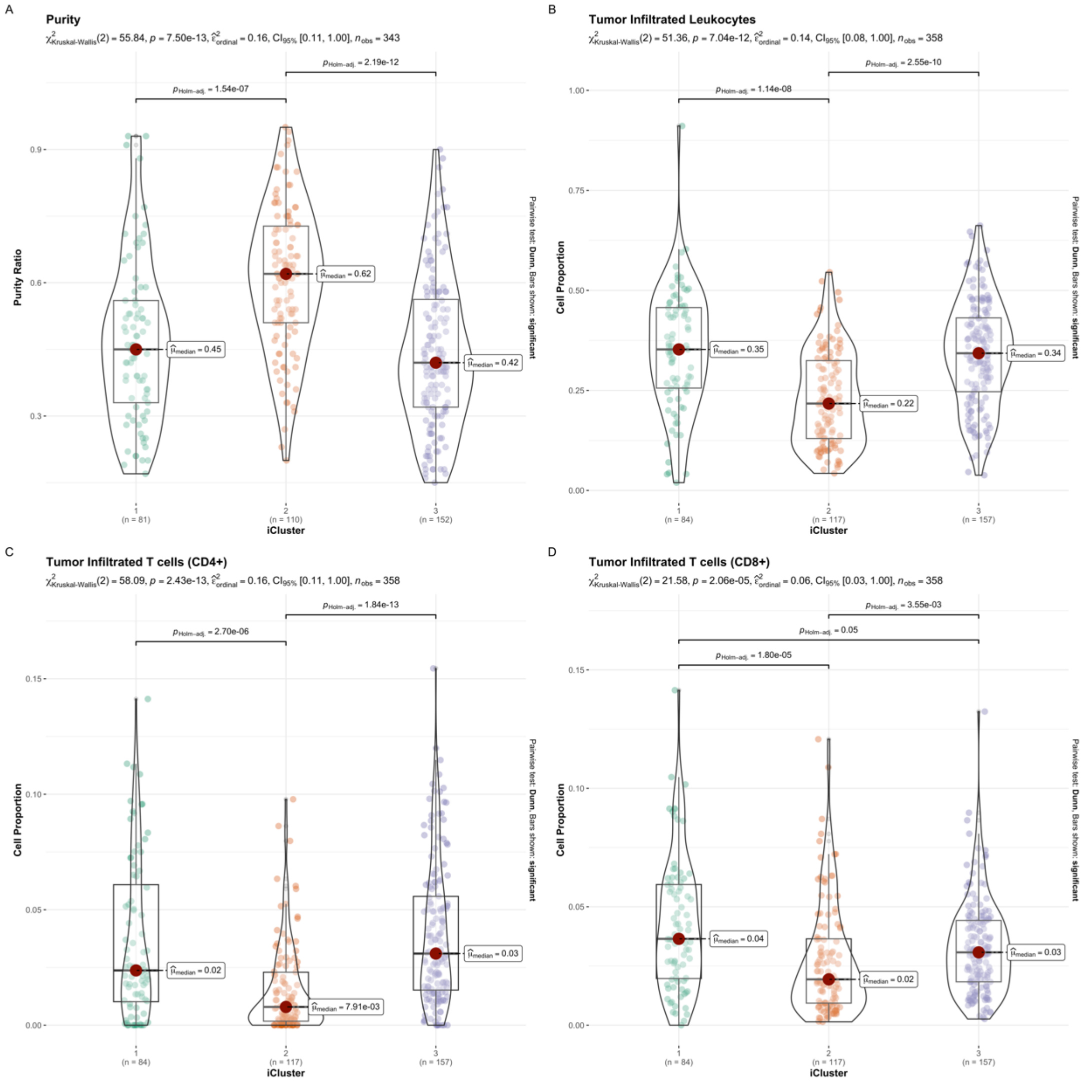


Figure 3.29. Comparison of LUSC iClusters for tumor infiltrated immune cells. (A) Purity ratio of tumor samples, (B) Fraction of tumor infiltrated leukocytes, (C) Proportion of CD4+ tumor infiltrated T cells, (D) Proportion of CD8+ tumor infiltrated T cells.

The differences of tumor mutation burden/load (TMB) and homologous DNA repair deficiency (HRD) level among iClusters shows significant results. iClust1 highest amount of TMB followed by iClust2 and iClust3 while iClust2 has higher HRD score than others (Figure 3.30).

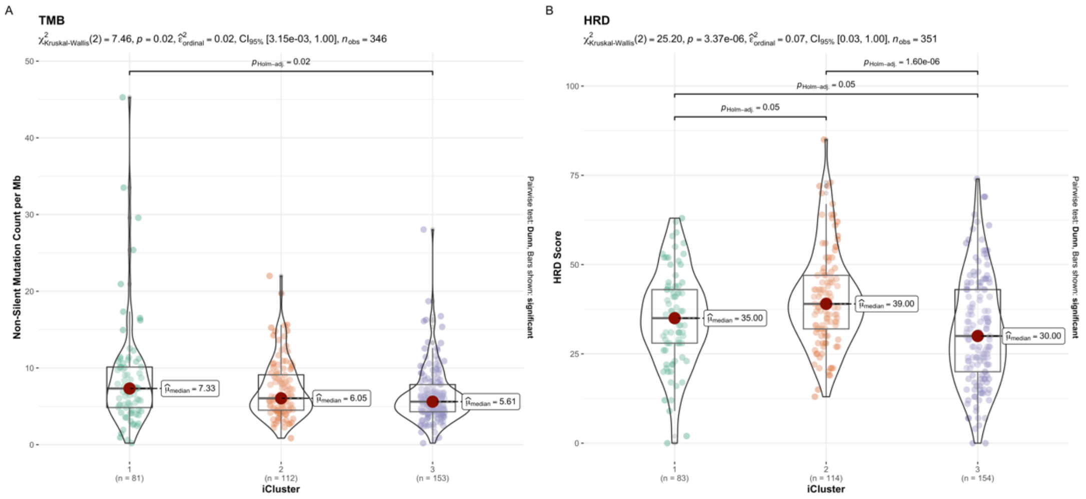


Figure 3.30. Comparison of LUSC iClusters for Tumor Mutation Burden (A) and Homologous DNA Repair Deficiency (B).

In order to see the correlation between TMB/HRD and tumor infiltrated immune cell fractions, we analyzed the correlation matrix independent of iClusters. TMB and HRD are positively correlated as expected. Fraction of tumor infiltrated leukocytes (TIL), CD4+ and CD8+ tumor infiltrated T cells are negatively correlated with HRD and TMB (Figure 3.31).

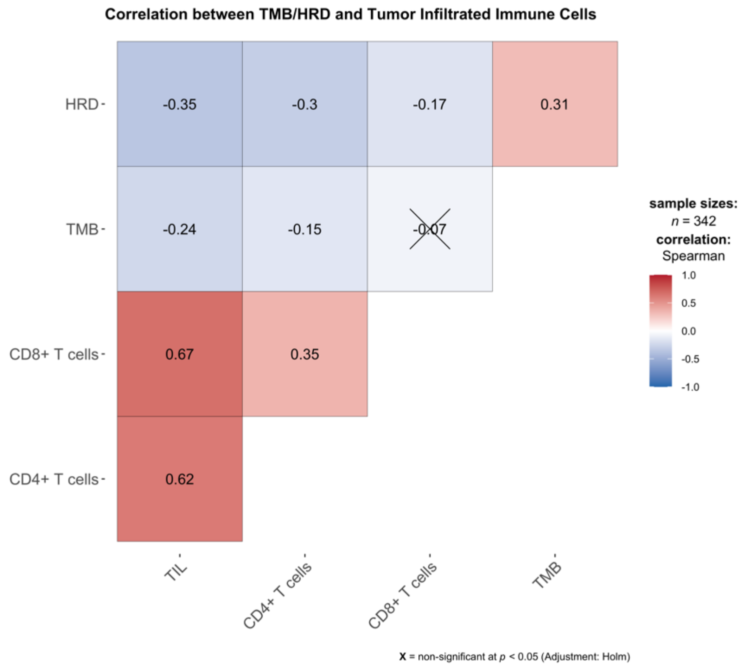


Figure 3.31. Correlation matrix between TMB/HRD and tumor infiltrated immune cells in LUSC tumor samples. TIL: Tumor Infiltrated Leukocytes, TMB: Tumor Mutation Burden, HRD: Homologous DNA Repair Deficiency.

We also analyzed expression differences of tumor associated proteins (TAPs) targetable by antibody therapeutics, among iClusters of LUSC patients. The expression pattern of TAPs is more diverse than pattern of LUAD tumor samples. The expression pattern of TAPs appears to correlate with iClusters and immune subtypes but not with risk groups (Figure 3.32).

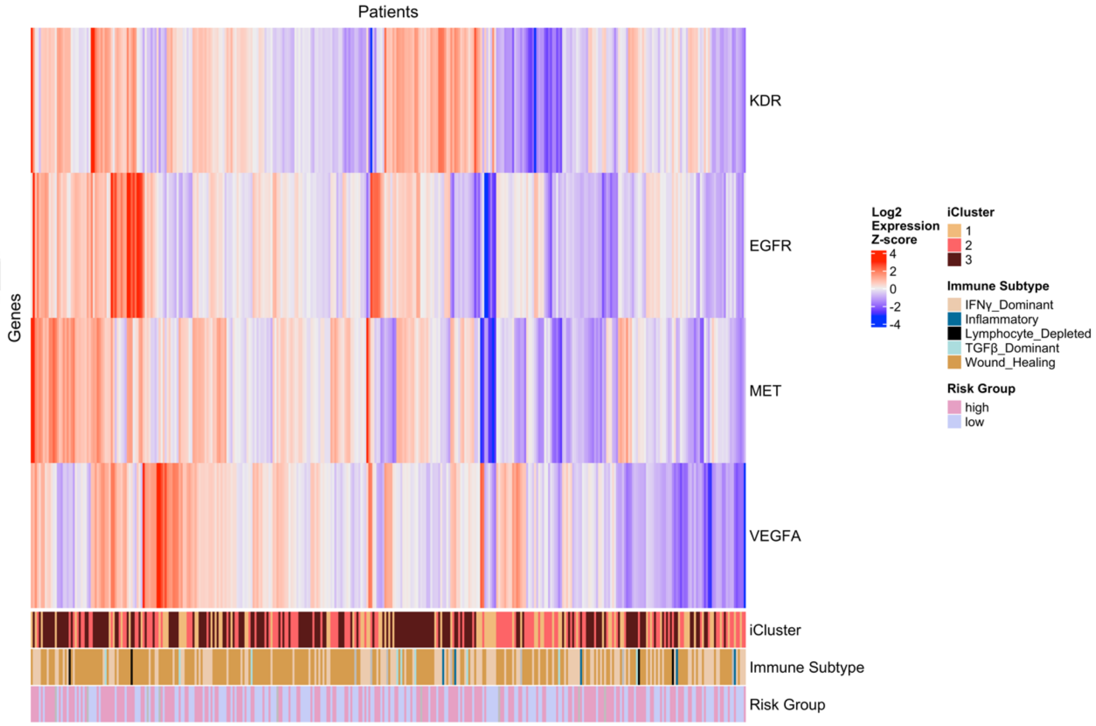


Figure 3.32. Expression heatmap of tumor associated proteins targetable by antibody therapeutics in LUSC tumor samples.

The expression of immune associated proteins (IAPs) including LAG3, CTLA4 and PD-L1 (CD274) which are targetable by antibody therapeutics, have obvious differences between tumor samples and patients may have different expression patterns of IAPs. While CTLA4 and LAG3 have close expression pattern, some patients have opposite expression of PD-L1 (CD274) and CTLA4 or LAG3. The expression pattern of IAPs appears to correlate with iClusters and immune subtypes but not with risk groups (Figure 3.33).

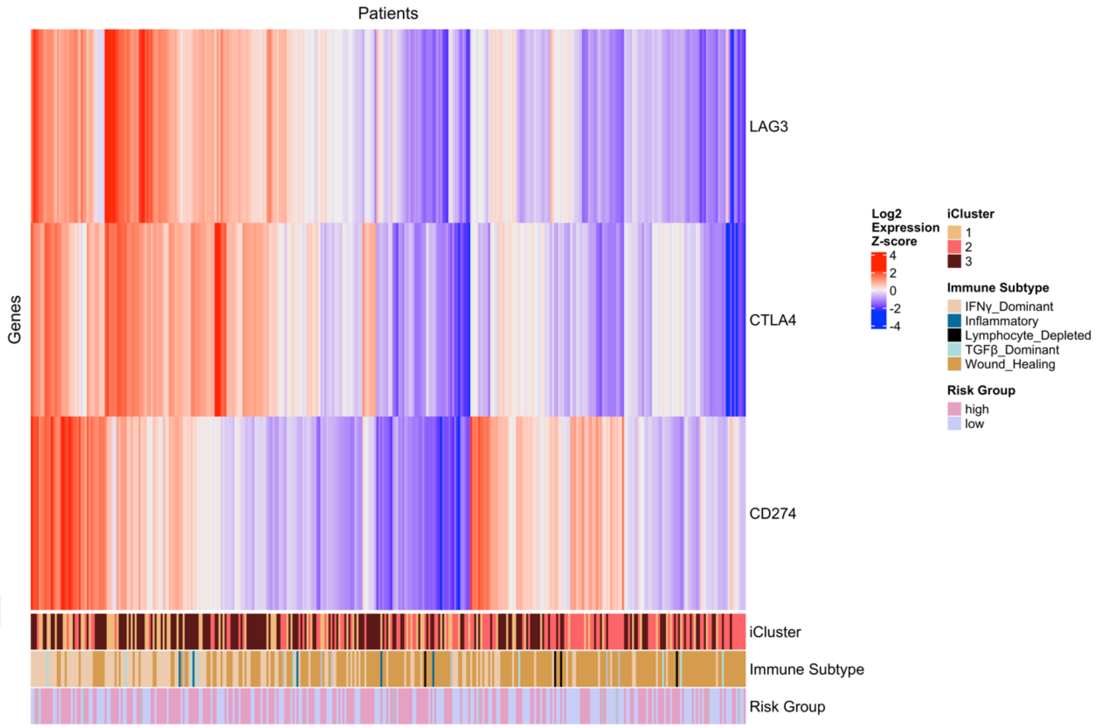


Figure 3.33. Expression heatmap of immune associated proteins targetable by antibody therapeutics in LUSC tumor samples.

We identified the relationships between iClusters and expression of TAPs and IAPs, using continuous variable comparison across iClusters. Violin plots show significant differences between iClusters for CTLA4, EGFR, KDR and MET genes. iCluster1 and iCluster3 have higher expression of CTLA4 and KDR while iCluster3 has highest expression of CTLA4, EGFR, KDR and MET (Figure 3.34).

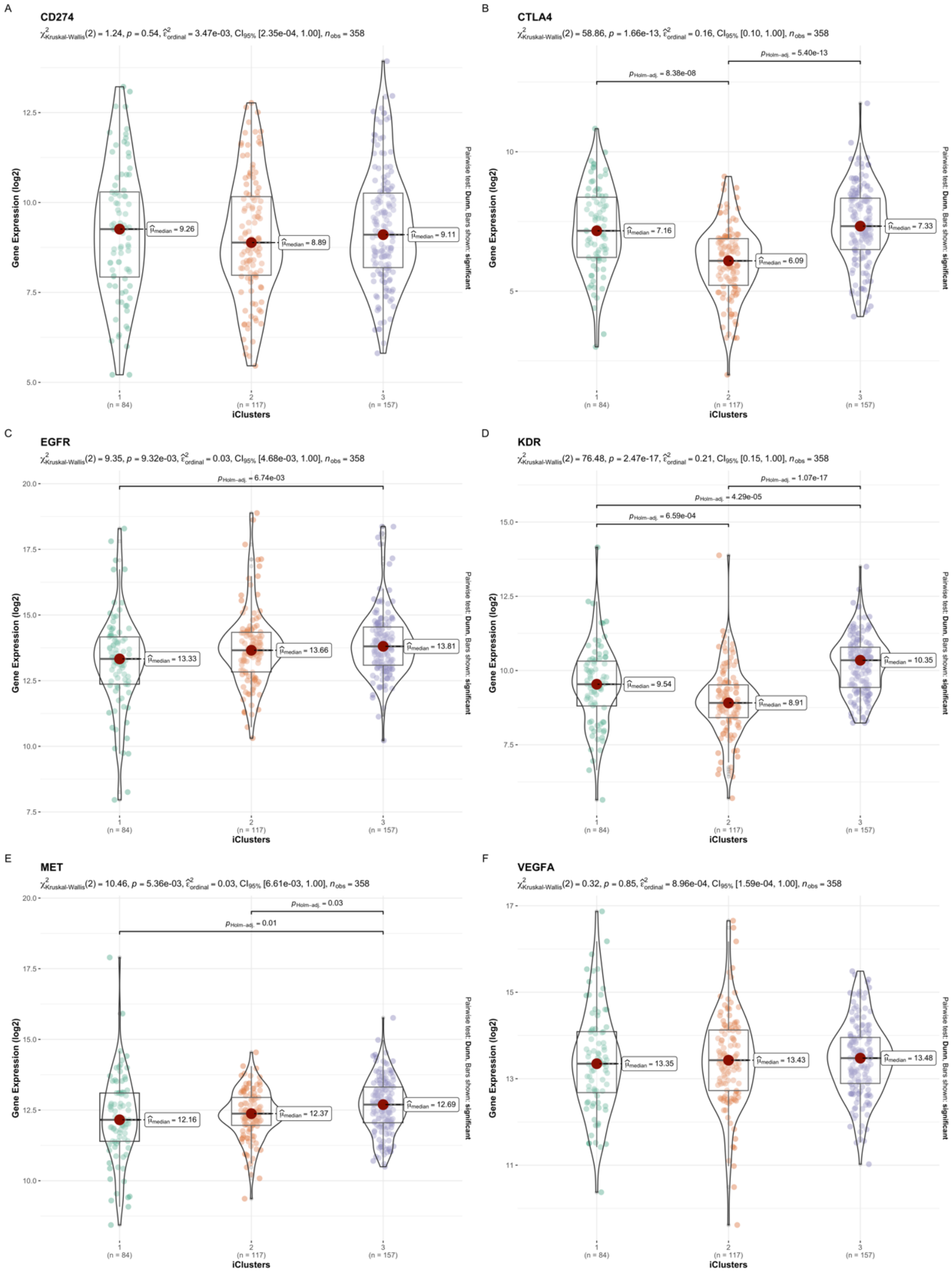


Figure 3.34. Comparison of LUSC iClusters for expression of targetable proteins by antibody therapeutics.

Risk clusters do not appear to be associated with expression of targetable proteins in heatmaps. Pearson's $\chi2$-test of independence showed that there is not a significant association between iClusters and risk clusters, but iCluser1 and iCluster3 have higher percentage of high-risk patients (Figure 3.35).

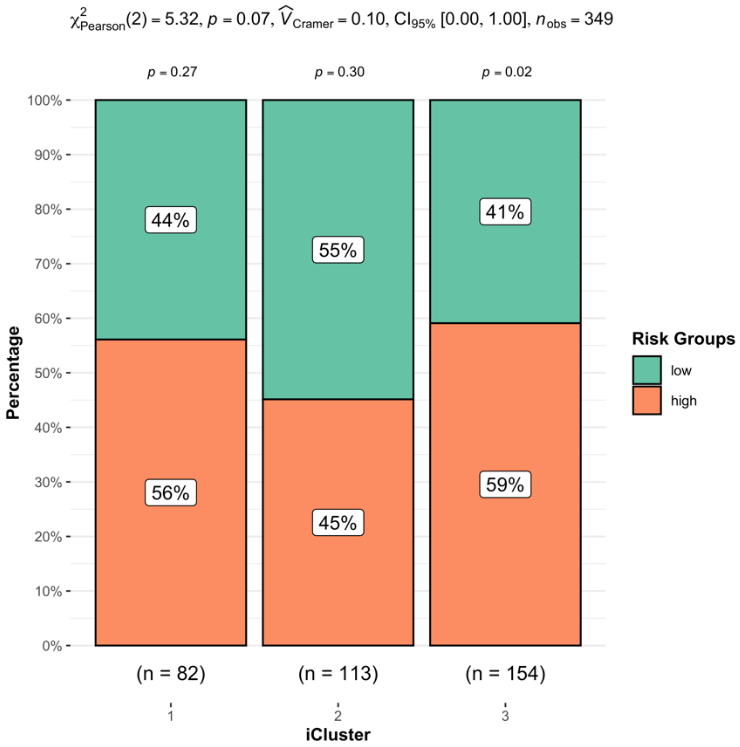


Figure 3.35. Relationships between iClusters and risk clusters of LUSC patients. Bar chart of iClusters vs risk clusters and results of Pearson’s $\chi 2$-test.

The expression levels of immunotherapeutic targetable genes among risk groups, show significant differences between risk groups for CTLA4, KDR and MET. Expression of CTLA4, KDR and MET is slightly lower in low-risk group (Figure 3.36).

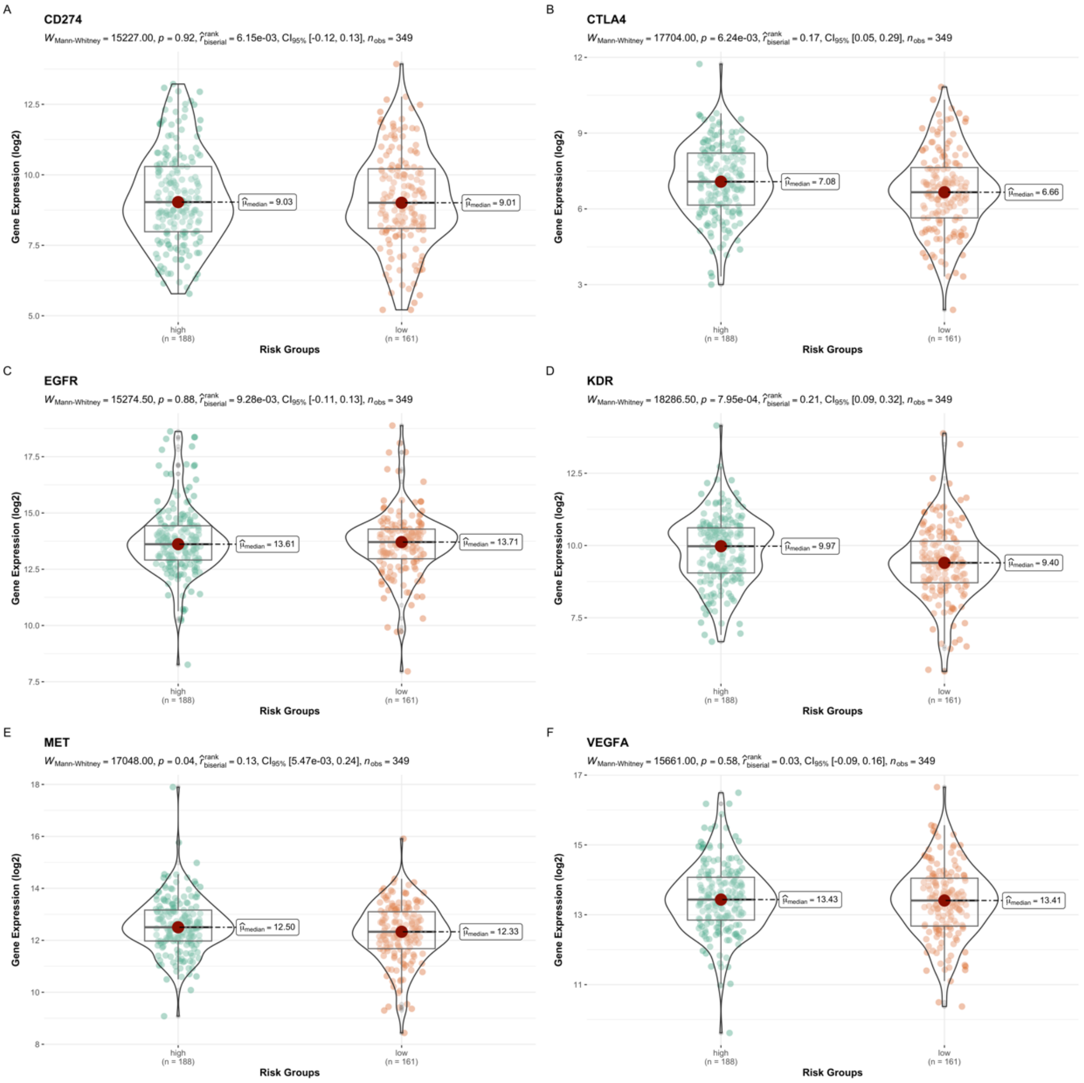


Figure 3.36. Comparison of LUSC risk clusters for expression of targetable proteins by antibody therapeutics.

3.4. Discussion

Using the TCGA data sets, we demonstrated the integrative clustering method, iClusterBayes, to identify clinically relevant patient clusters and driver data features. Integrative clustering of the LUAD patients using multiple data types (SNVs, CNVs, RNAseq, miRNAseq and methylation) revealed four integrative clusters (iClusters) with distinct genomic/transcriptomic patterns. The LUAD patients in the iCluster2 exhibited a more favorable survival outcome compared to those in the other three clusters. The driver genes identified in the study, such as TP53, KEAP1, STK11, KRAS and EGFR, are known to be associated with the disease. These genes exhibited varying mutation frequencies across the four iClusters. Among them, tumor suppressor genes TP53 and STK11, as well as oncogenes KRAS and EGFR, were identified as driver genes with distinct mutation patterns in the different iClusters. Targeting these oncogenes may hold potential therapeutic value for patients within specific iCluster subgroups displaying abnormal expression or mutations. Additionally, the expression patterns of immune response, PI3K-Akt signaling, and cell adhesion pathways differed among the iClusters, indicating their potential as therapeutic targets as well.

Integrative clusters (iClusters) have an association with immune subtypes of TCGA cancer samples (Thorsson et al., 2018). iCluster2 has the highest percentage of inflammatory subtype which has better prognosis among others while iCluster1 has the lowest percent of inflammatory subtype and higher amount of wound healing and IFN-γ dominant subtype which have worse prognosis. iCluster2 has less purity with higher fraction of tumor infiltrated leukocytes (TILs) and CD4+ T cells. Besides, it has less amount of tumor mutation burden (TMB) and homologous DNA repair deficiency (HRD) score. On the other hand, iCluster1 and iCluster4 have higher amount of TMB and HRD while they have higher purity with less fraction of TILs and CD4+ T cells. iCluster2 has higher expression of all targetable tumor associated proteins (TAPs) and tumor associated immune proteins (IAPs) while there is not any difference in expression of targetable proteins except CTLA4 among risk clusters.

We identified three distinct iClusters for LUSC (lung squamous cell carcinoma) based on their mutation, copy number variation, mRNA, and miRNA expression patterns. These iClusters also had clinical significance as their overall survival rates exhibited significant differences. In terms of expression subtypes, the secretory and classical subtypes are like the iCluster1 and iCluster2, respectively. Basal cell carcinoma and Wnt signaling pathways showed different expression pattern between the three iClusters, which have possibility of utilizing these characteristics as targets for cluster specific therapies. There is no association between immune subtypes and iClusters of

LUSC. Moreover, immune clusters don't have significant survival differences in LUSC. However, iClusters have significant difference: iCluster1 has poorer prognosis than others. iCluster1 and iCluster3 have higher expression of CTLA4 and KDR significantly while expression of CTLA4, KDR and MET is slightly lower in low-risk group. iCluster1 and iCluster3 have less purity with higher fraction of TILs, CD4+ T cells, CD8+ T cells and less HRD. iCluster2 has the opposite pattern for all variables. iCluster1 has higher TMB than iCluster3 although HRD and TMB have positive correlation, and iCluster1 has worse prognosis than iCluster3. When examining the findings of iClusters for LUAD (lung adenocarcinoma) and LUSC (lung squamous cell carcinoma), it is observed that tumor mutation burden (TMB) shows a negative correlation with survival probability. In LUSC, tumor-infiltrating immune cells exhibit a negative correlation with TMB and homologous recombination deficiency (HRD), while they show a positive correlation with survival in both LUAD and LUSC. In LUAD, TMB and HRD display a negative correlation with CD4+ T cells but a positive correlation with CD8+ T cells.

In Immune Landscape of Cancer paper, it was found that the lymphocyte fraction (LF) positively correlates with tumor mutation burden (TMB), loss of heterozygosity (LOH), aneuploidy and homologous recombination deficiency (HRD), but negatively correlates with copy number variation (CNV) burden (Thorsson et al., 2019). Furthermore, tumors with higher TMB, higher number of mutations in DNA repair proteins and higher neoantigen burden, show improved response and progression-free survival, particularly in response to anti-PD-1 therapy inducing neoantigen-specific T cell activity (Rizvi et al., 2015). We observed conflicting results that tumor infiltrated lymphocyte fraction and CD4+ T cells have negative correlation with TMB/HRD in both LUAD and LUSC. CD8+ T cells have positive correlation with TMB/HRD in LUAD but negative correlation in LUSC. However, independent from these correlations, higher amount of TIL and CD4+/CD8+ T cells showed better prognosis in both LUAD and LUSC. Previously published researches showed that higher ratio of "pre-exhausted" T cells is associated with a better prognosis (Guo et al., 2018) and additionally, higher ratio of CD8+ cytotoxic T cells is significantly linked to better progression-free interval (PFI) and overall survival (OS), especially in patients with programmed death ligand-1 (PD-L1) positivity and treated with immune checkpoint inhibitors (Li et al., 2021; Lopez de Rodas et al., 2022). iClusters which are parallel with these findings may help to predict the outcome or response of ICI or other antibody therapies. However, we need to perform further analyses such as iCluster based correlations to see the relationship in existence or absence of TILs in tumor sample. Leader et al. identified a cellular module called Lung Cancer Immune Activation Module consisting higher amount of TMB, higher expression of cancer testis antigen and TP53mut with PDCD1+, CXCL13+ activated T cells, IgG+ plasma cells and SPP1+ macrophages referred to activation module (LCAM-hi). LCAM-hi is correlated with superior anti-PDL1 response while LCAM-lo consisting lower amount of TMB, lower expression of cancer testis antigen and TP53WT with naive T cells, lower plasma:B cell ratio, tissue-resident macrophages and resting DCs is correlated with worse anti-PDL1 response (Leader et al., 2021).

Expression of CTLA4 is correlated with fraction of TILs in LUAD clusters. LUAD iCluster2 may be best candidate group of patients who may give better response to immune check point inhibitors because iCluster2 patients have higher expression of both CTLA4 and CD274 (PDL1). Expression of CTLA4 is correlated with fraction of TILs in LUSC clusters, too. LUAD and LUSC datasets do not have response data, therefore we could not compare them for the responses, however we plan to analyze iClusters and tumor/immune biomarkers using cancer type containing response data.

In summary, we have performed iClusterBayes method for integrative clustering patient samples separately in 33 TCGA projects however we presented only results of LUAD and LUSC for the consistency in the story of the thesis. This approach offers researchers a robust tool for deciphering cancer omics data, enabling the identification of clinically significant cancer subtypes and potential markers for therapeutic interventions. However, it is necessary to test the method several times with independent data sets and to perform trials before making clinical decisions.

4. CHAPTER

TCGAnalyzeR: A Web Application for Integrative Visualization of Molecular and Clinical Data of Cancer Patients for Cohort and Associated Gene Discovery[2]

4.1. Abstract

The Cancer Genome Atlas (TCGA) database, which contains comprehensive molecular and clinical data on over 11,000 cancer patients with 33 different cancer types, is incredibly large and complicated, making it difficult to use this important resource effectively. We present TCGAnalyzeR, a web tool that integrates visualization of pre-processed TCGA data with numerous modules: (i) Simple nucleotide variations with driver prediction, pathway enrichment and survival analysis; (ii) Copy number variations with pathway enrichment and survival analysis; (iii) Differential expression in tumors versus normal with pathway enrichment and survival analysis; (iv) Clinical data with survival analysis, and descriptive graphics; (v) Internal patient clusters generated using the iClusterPlus R package or signature-based expression analysis; (vi) Subtypes from the literature, curatedTCGAData and BiocOncoTK R packages. TCGAnalyzeR offers cancer researchers dynamic, integrated representations of this multi-omic, pan-cancer TCGA data, along with displaying cohort- or gene-centric results. Users can design their own custom gene sets for pan-cancer comparisons, custom patient subcohorts to compare external subtypes (MSI, Immune, PAM50, Triple Negative, IDH1, miRNA, etc.) and our internal patient clusters.

---

[2] **TCGAnalyzeR: A Web Application for Integrative Visualization of Molecular and Clinical Data of Cancer Patients for Cohort and Associated Gene Discovery**
Talip Zengin, Başak Abak Masud, Tuğba Önal-Süzek. bioRxiv, 2023, doi: 10.1101/2023.01.20.524925

4.2. Methods

4.2.1. Data

Publicly available hg38 data including SNV (mutation annotation format [maf] files generated by mutect2 pipeline), CNV, Transcriptome Profiling (normalized HTseq counts) and clinical data of 33 cancer types from The Cancer Genome Atlas (TCGA) projects were downloaded on March 6, 2022, from NCI GDC (Grossman et al., 2016) using TCGAbiolinks R package (Colaprico et al., 2016).

4.2.2. SNV Analysis

Potential driver mutated genes with their roles as a tumor suppressor or oncogene were determined by SomInaClust R package (Van den Eynden et al., 2015). The COSMIC Mutation Data and Cancer Gene Census data were used for mutation validation from Catalog of Somatic Mutations in Cancer (COSMIC), organized by expert scientists with a careful review of numerous scientific publications (https://cancer.sanger.ac.uk/cosmic) (Forbes et al., 2017).

4.2.3. CNV Analysis

Significant recurrent copy number variations were identified by gaia R package (Morganella et al., 2011). NCBI IDs and Hugo Symbols of the genes on chromosomal regions were determined using GenomicRanges (Lawrence et al., 2013) and biomaRt (Durinck et al., 2009) packages.

4.2.4. Differential Expression Analysis (DEA)

Differentially expressed genes were determined by limma-voom method using edgeR (Robinson et al., 2010) and limma (Ritchie et al., 2015) packages. Ensembl IDs were converted to NCBI IDs and Hugo Symbols using the biomaRt package (Durinck et al., 2009). If it is available for a particular cancer, two different DEA were performed using two datasets: (i) *Paired*: Tumor samples against tumor-adjacent normal samples; (ii) *All*: Tumor samples of all patients against normal samples of patients.

Pathway enrichment and visualization was performed for each analysis by clusterProfiler R package (Yu et al., 2012).

4.2.5. Pre-computed Patient Clusters and Sample Subtypes

TCGAnalyzeR provides an interactive visual analysis of several patient cohorts: i) Risk Clusters: Low-risk or high-risk patient clusters determined by expression-based gene signature analysis for Lung Adenocarcinoma (LUAD), Lung Squamous Cell Carcinoma (LUSC) and Colon Adenocarcinoma (COAD) (Zengin and Önal-Süzek, 2021), ii) iClusters: Integrative patient clusters using raw SNV, CNV, gene expression, miRNA expression and methylation data of tumor samples by iClusterBayes method (Mo et al., 2018), iii) Curated subtypes (immune subtypes, TNBC subtypes, PAM50 subtypes) from original publications (Thorsson et al, 2018; Lehmann et al, 2016; Berger et al, 2018), previously published TCGA subtypes from curatedTCGA R package (Ramos et al., 2020), sample subtypes based on Microsatellite Instability (MSI) from BiocOncoTK (Carey et al., 2022, Ding et al., 2018) and Immune subtypes (Thorsson et al., 2019) for available cancer types. In total, 123 external patient cohorts are integrated into the web interface allowing efficient filtering and cross-comparative analysis of multiple subtypes in parallel.

4.2.6. Survival Analysis

Kaplan-Meier (KM) survival analysis is performed using *survfit* R package (Therneau et al., 2022) in real-time based on reading overall survival data of patients of interest for selected clinical features by *readr* R package (Wickham et al., 2022).

4.2.7. Visualization

TCGAnalyzeR front-end was implemented by javascript-based R packages with an interactive dashboard enabling users to select cancer types, data types, risk groups and patient cohorts using *heatmaply*, *g3viz* and *highcharter* R packages (Galili et al., 2018; Guo et al., 2019; Kunst et al., 2022). All visualizations are interactive and customizable by the user through the filtration options with “My genes” and/or “My patients” panels enabling to copy genes and/or patients of interest to the clipboard.

All methods used for TCGAnalyzeR web application are summarized in Figure 4.1.

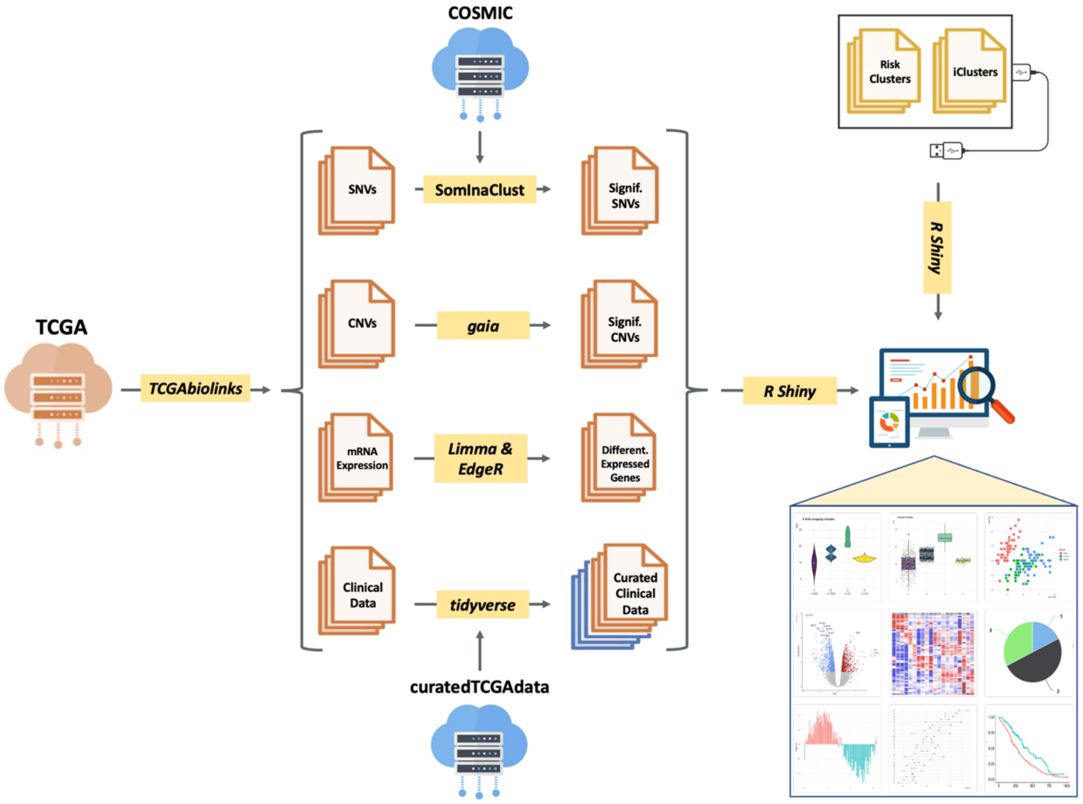


Figure 4.1. Flowchart of methods used for preparing the data for TCGAnalyzeR web tool.

4.3. Examples of Data Visualization from TCGAnalyzeR

Users can find summaries of the TCGA cancer datasets and selected cancer type at main page of the tool such as bar graphs showing numbers of samples and genes.

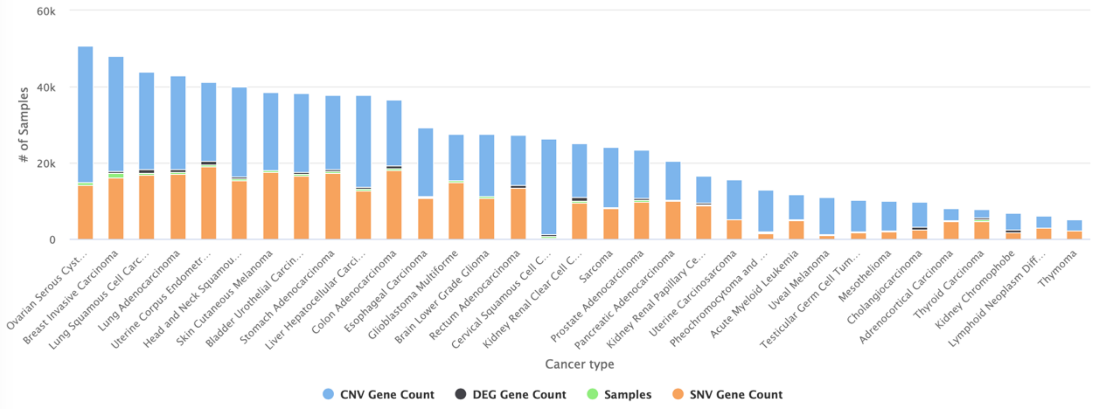


Figure 4.2. Bar graph summarizing numbers of samples, genes with SNVs, genes with CNVs and Differentially Expressed Genes (DEGs) at main page of TCGAnalyzeR web tool.

As example we present the summary and visualization of data analysis results for TCGA lung adenocarcinoma (LUAD) dataset below.

Mutations of top 10 genes with somatic simple nucleotide variations (SNVs) predicted as driver genes by SomInaClust are shown in oncoplot with iClusters as bottom annotation (Figure 4.3). All genes are mutated in all clusters although some patients do not have a particular gene mutation or there are some patients who don't have mutations on any of these genes. iCluster1 has less patients with KRAS mutation, iCluster2 and iCluster3 have less patients with KEAP1 mutation and iCluster2 has less patients with CDH10 mutation as differences between clusters.

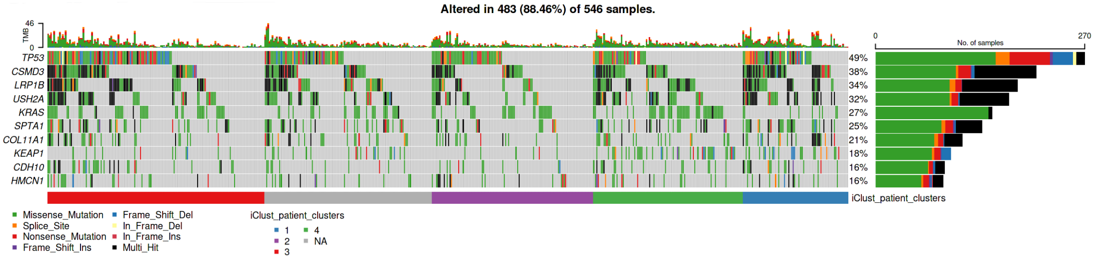
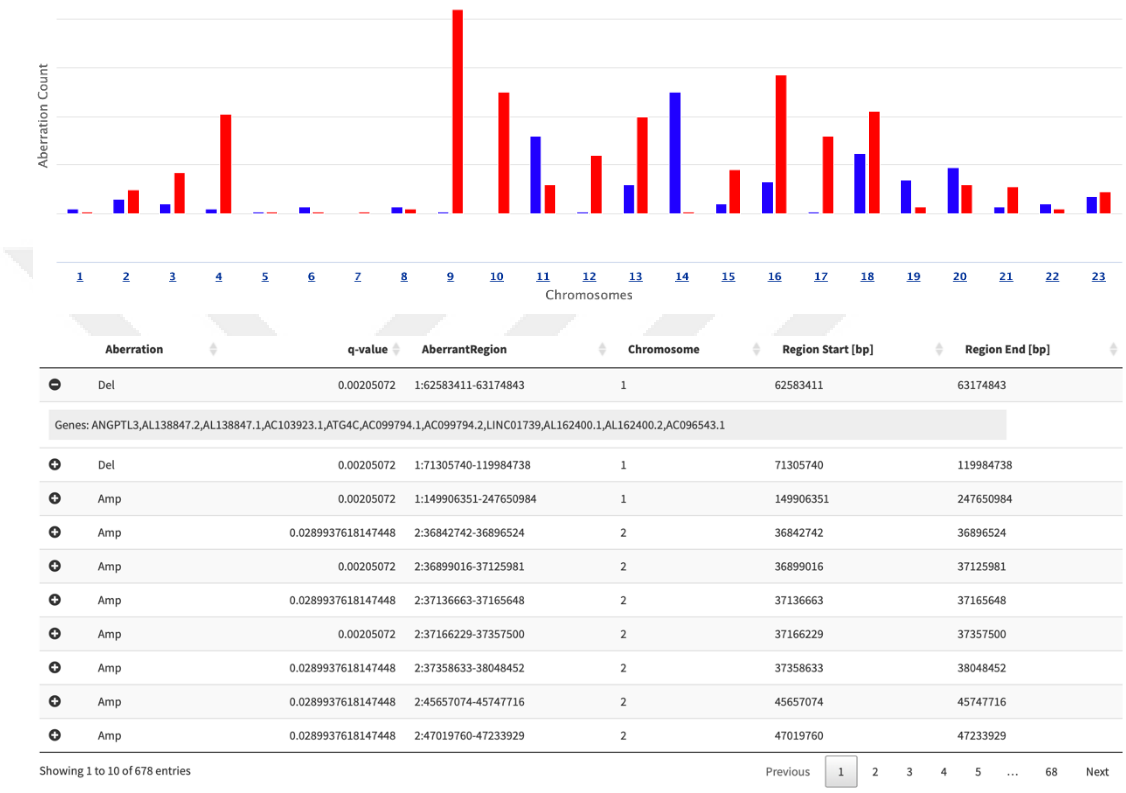


Figure 4.3. Oncoplot of genes with SNVs predicted as driver by SomInaClust.

Genomic regions with copy number variations (CNVs) are shown as bar graph separately based on their chromosomes and table with q-values of CNV analysis. According to results below, chromosome 9 has the highest number of amplifications while chromosome 14 has the highest number of deletions. Chromosomes 1, 5, 6, 7, 8 and 22 have very less aberration than other chromosomes (Figure 4.4).



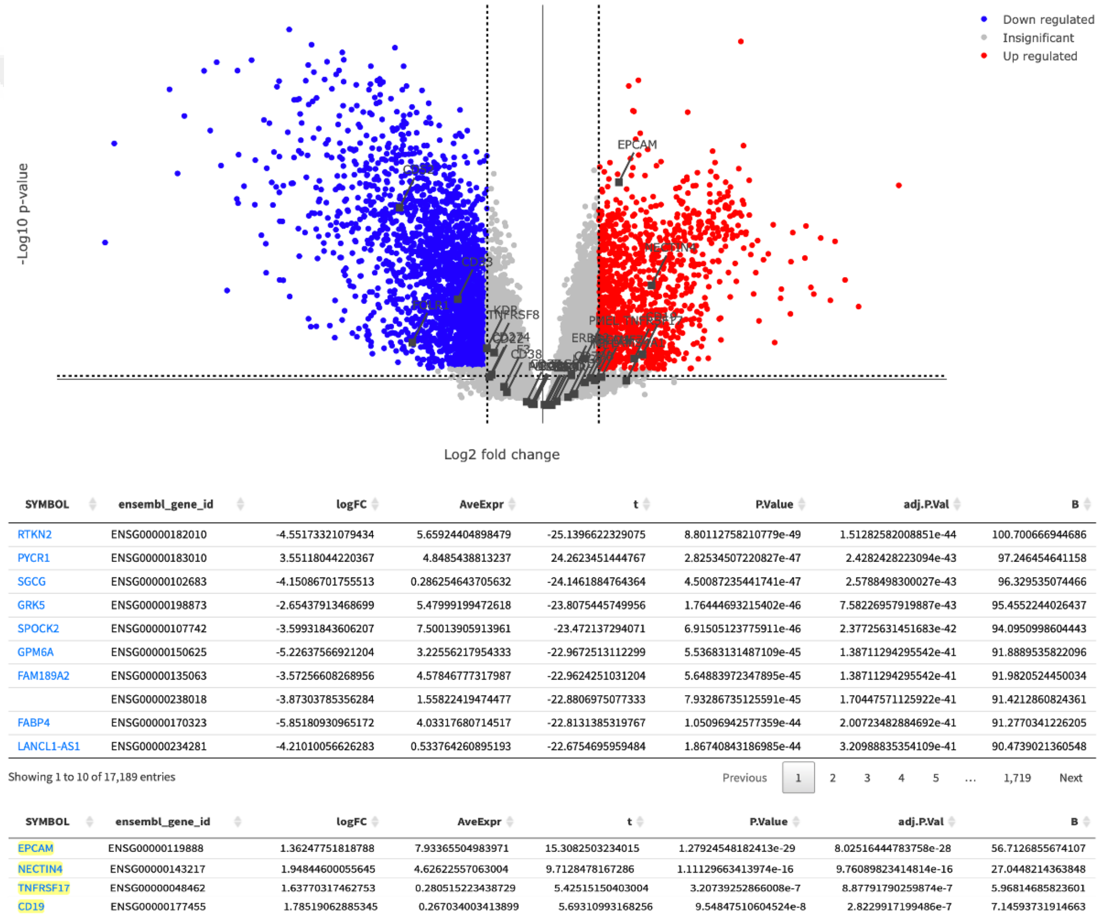
	Aberration	q-value	AberrantRegion	Chromosome	Region Start [bp]	Region End [bp]
⊖	Del	0.00205072	1:62583411-63174843	1	62583411	63174843
Genes: ANGPTL3,AL138847.2,AL138847.1,AC103923.1,ATG4C,AC099794.1,AC099794.2,LINC01739,AL162400.1,AL162400.2,AC096543.1						
⊕	Del	0.00205072	1:71305740-119984738	1	71305740	119984738
⊕	Amp	0.00205072	1:149906351-247650984	1	149906351	247650984
⊕	Amp	0.0289937618147448	2:36842742-36896524	2	36842742	36896524
⊕	Amp	0.00205072	2:36899016-37125981	2	36899016	37125981
⊕	Amp	0.0289937618147448	2:37136663-37165648	2	37136663	37165648
⊕	Amp	0.00205072	2:37166229-37357500	2	37166229	37357500
⊕	Amp	0.0289937618147448	2:37358633-38048452	2	37358633	38048452
⊕	Amp	0.0289937618147448	2:45657074-45747716	2	45657074	45747716
⊕	Amp	0.0289937618147448	2:47019760-47233929	2	47019760	47233929

Showing 1 to 10 of 678 entries

Figure 4.4. Bar graph and CNV analysis results of genomic regions with CNVs in LUAD dataset. Amp/Red: Amplification, Del/Blue: Deletion

Differentially expressed genes in tumor samples versus adjacent normal samples are shown as volcano plot and in summary table with log2 fold change (logFC) and statistics results in Figure 4.5. Genes of all targetable proteins by antibody therapeutics are highlighted in volcano plot. Only EPCAM, NECTIN4, TNFRSF17 and CD19 genes are significantly up-regulated in LUAD tumor samples. EPCAM gene encodes the protein called Epithelial Cell Adhesion Molecule which is targeted by edrecolomab in colorectal cancer and catumaxomab in ovarian cancer, malignant ascites and gastric cancer. NECTIN4 encodes Nectin Cell Adhesion

Molecule 4 protein targeted by enfortumab vedotin in urogenital cancer. TNFRSF17 encodes Tumor necrosis factor receptor superfamily member 17 protein, also known as B cell maturation antigen (BCMA), which plays role in the regulation of B cell survival and immunity and targeted by belantamab mafodotin and teclistamab in multiple myeloma. CD19 encodes B-lymphocyte antigen CD19 which is expressed in B cells and targetable by tafasitamab, loncastuximab tesirine and blinatumomab in B cell lymphomas. The up-regulation of TNFRSF17 and CD19 may depend on the infiltration of B cells in tumor samples.



SYMBOL	ensembl_gene_id	logFC	AveExpr	t	P.Value	adj.P.Val	B
RTKN2	ENSG00000182010	-4.55173321079434	5.65924404898479	-25.1396622329075	8.80112758210779e-49	1.51282582008851e-44	100.700666944686
PYCR1	ENSG00000183010	3.55118044220367	4.8485438813237	24.2623451444767	2.82534507220827e-47	2.4282428223094e-43	97.246454641158
SGCG	ENSG00000102683	-4.15086701755513	0.286254643705632	-24.1461884764364	4.50087235441741e-47	2.5788498300027e-43	96.329535074466
GRK5	ENSG00000198873	-2.65437913468699	5.47999199472618	-23.8075445749956	1.76444693215402e-46	7.58226957919887e-43	95.4552244026437
SPOCK2	ENSG00000107742	-3.59931843606207	7.50013905913961	-23.472137294071	6.91505123775911e-46	2.37725631451683e-42	94.0950998604443
GPM6A	ENSG00000150625	-5.22637566921204	3.22556217954333	-22.9672513112299	5.53683131487109e-45	1.38711294295542e-41	91.8889535822096
FAM189A2	ENSG00000135063	-3.57256608268956	4.57846777317987	-22.9624251031204	5.64883972347895e-45	1.38711294295542e-41	91.9820524450034
	ENSG00000238018	-3.87303785356284	1.55822419474477	-22.8806975077333	7.93286735125591e-45	1.70447571125922e-41	91.4212860824361
FABP4	ENSG00000170323	-5.85180930965172	4.03317680714517	-22.8131385319767	1.05096942577359e-44	2.00723482884692e-41	91.2770341226205
LANCL1-AS1	ENSG00000234281	-4.21010056626283	0.533764260895193	-22.6754695959484	1.86740843186985e-44	3.20988835354109e-41	90.4739021360548

Showing 1 to 10 of 17,189 entries

Previous 1 2 3 4 5 … 1,719 Next

SYMBOL	ensembl_gene_id	logFC	AveExpr	t	P.Value	adj.P.Val	B
EPCAM	ENSG00000119888	1.36247751818788	7.93365504983971	15.3082503234015	1.27924548182413e-29	8.02516444783758e-28	56.7126855674107
NECTIN4	ENSG00000143217	1.94844600055645	4.62622557063004	9.7128478167286	1.11129663413974e-16	9.76089823414814e-16	27.0448214363848
TNFRSF17	ENSG00000048462	1.63770317462753	0.280515223438729	5.42515150403004	3.20739252866008e-7	8.87791790259874e-7	5.96814685823601
CD19	ENSG00000177455	1.78519062885345	0.267034003413899	5.69310993168256	9.54847510604524e-8	2.8229917199486e-7	7.14593731914663

Figure 4.5. Volcano plot and analysis results of differentially expressed genes of LUAD samples. Genes of all targetable proteins by approved antibody therapeutics are highlighted in volcano plot. Statistics of significantly up-regulated genes are summarized in the second table.

Pathway enrichment results of differentially expressed genes (DEGs) are summarized as bar graph in Figure 4.6. DEGs mostly play role in extra cellular pathways such as cell adhesion and ECM-receptor interaction; immune related pathways including complement and coagulation cascades, hematopoietic cell lineage, infection and cytokine-cytokine receptor interaction; cell cycle and protein digestion and absorption pathways (Figure 4.6).

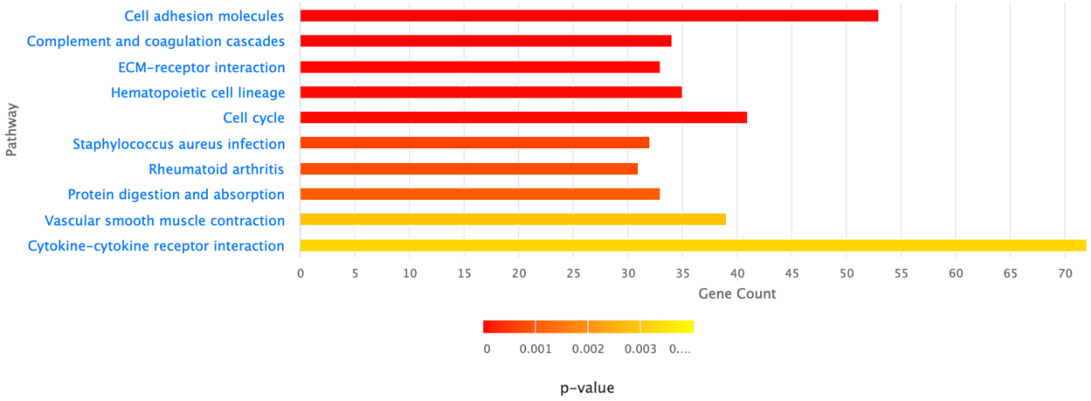


Figure 4.6. Pathway enrichment of differentially expressed genes in LUAD tumor samples.

If user wants to visualize the expression level of a specific gene in tumor samples versus adjacent normal cells, violin plot can be generated with adjusted p-value (q-value) calculated during differential expression analysis. According to violin plot shown in Figure 4.7, EPCAM gene is up-regulated highly significantly in tumor samples.

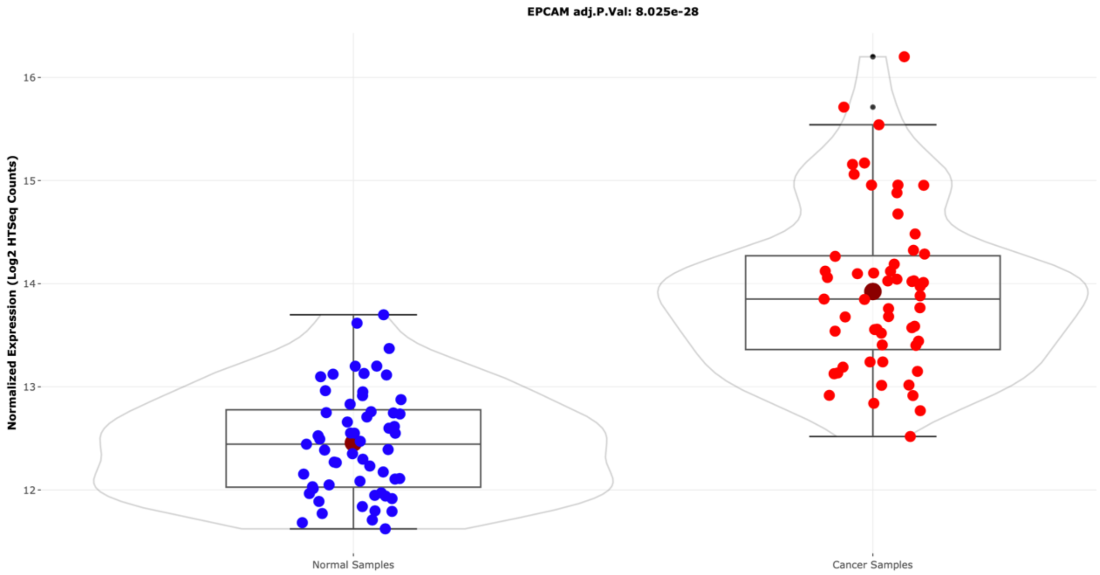
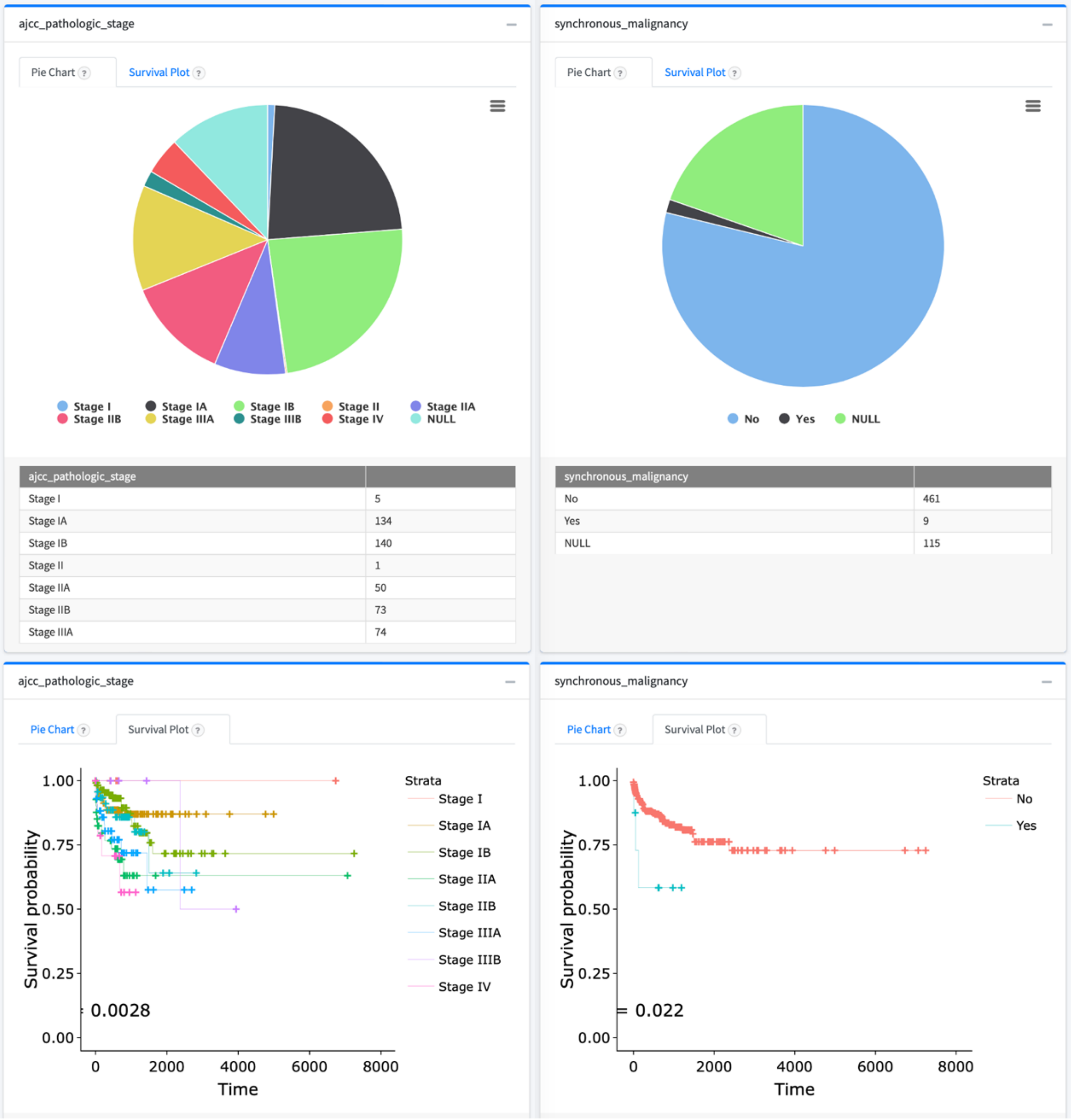


Figure 4.7. EPCAM gene expression in tumor and adjacent normal samples of LUAD.

We can summarize the clinical variables as pie charts and survival analysis can performed in real-time and present in other tab. We can see the distribution of tumor stage (ajcc pathologic stage) categories and existence of synchronous malignancy as pie chart with survival analysis results in Figure 4.8. Survival probabilities of tumor stages are significantly different and patients who have synchronous malignancy have poor prognosis significantly.



ajcc_pathologic_stage	
Stage I	5
Stage IA	134
Stage IB	140
Stage II	1
Stage IIA	50
Stage IIB	73
Stage IIIA	74

synchronous_malignancy	
No	461
Yes	9
NULL	115

Figure 4.8. Pie charts and survival analysis of pathologic stage and existence of synchronous malignancy. There are many clinical variables in clinical data section of LUAD dataset but only two of them is showed in this figure.

4.4. Conclusion

We present comprehensive analyses of genetic mutations, copy number variations, and differential gene expression in large sets of patient clusters with subtype information. Our approach includes signature-based clustering using the Generalized Linear Model for two specific cancer types (LUAD and LUSC). Additionally, we provide immune and MSI-sensor scores for all 33 cancer types, as well as subtype information for breast cancer (BRCA) including PAM50 and TNBC patient cohorts, retrieved from previously published TCGA articles. For fifteen cancer types, TCGA subtype information was retrieved the curatedTCGAData R package (Ramos et al, 2020).

TCGAnalyzeR is a user-friendly web tool designed for integrated and large-scale analyses of genomic and clinical data from TCGA across 33 cancer types. Through the TCGAnalyzeR web interface, cancer researchers can easily create subcohorts and gene sets of interest to filter and visualize the results. The TCGAnalyzeR help page provides a demonstration of the tool, showcasing two specific use-cases for subcohort discovery.

The full version of the article can be found at Appendix B.

REFERENCES

Angell H, Galon J. From the immune contexture to the Immunoscore: the role of prognostic and predictive immune markers in cancer. Curr Opin Immunol. 2013 Apr; 25(2):261-7.

Aran, D., Hu, Z. & Butte, A. J. XCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol. 2017, 18, 220.

Bai, R., Lv, Z., Xu, D. et al. Predictive biomarkers for cancer immunotherapy with immune checkpoint inhibitors. Biomark Res. 2020, 8, 34.

Bao X, Shi R, Zhao T, Wang Y. Immune landscape and a novel immunotherapy-related gene signature associated with clinical outcome in early-stage lung adenocarcinoma. J Mol Med (Berl). 2020 Jun;98(6):805-818.

Beck B, Blanpain C. Unraveling cancer stem cell potential. Nat Rev Cancer 2013;13(10):727–38.

Bindea, G. et al. Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. Immunity 2013, 39, 782–795.

Binnewies, M., Roberts, E.W., Kersten, K. Understanding the tumor immune microenvironment (TIME) for effective therapy. Nat Med 2018, 24, 541–550.

Brahmer J, Reckamp KL, Baas P. Nivolumab versus Docetaxel in advanced squamous-cell non-small-cell lung Cancer [J]. N Engl J Med. 2015;373(2):123–35.

Cao R, Yuan L, Ma B, Wang G, Tian Y. Tumour microenvironment (TME) characterization identified prognosis and immunotherapy response in muscle-invasive bladder cancer (MIBC). Cancer Immunol Immunother. 2021 Jan;70(1):1-18.

Cao T, Shen H. Development of a multi-gene-based immune prognostic signature in ovarian Cancer. J Ovarian Res. 2021 Jan 28;14(1):20.

Chen, H. VennDiagram: Generate High-Resolution Venn and Euler Plots. R Package Version 1.6.20. 2018. Available online: https://cran.r-project.org/package=VennDiagram (accessed on 21 May 2020).

Cheng J, Ding X, Xu S, Zhu B, Jia Q. Gene expression profiling identified TP53MutPIK3CAWild as a potential biomarker for patients with triple-negative breast cancer treated with immune checkpoint inhibitors. Oncol Lett. 2020 Apr;19(4):2817-2824.

Colaprico, A.; Silva, T.C.; Olsen, C.; Garofano, L.; Cava, C.; Garolini, D.; Sabedot, T.S.; Malta, T.M.; Pagnotta, S.M.; Castiglioni, I.; et al. TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 2016, 44, e71.

Durinck, S.; Spellman, P.T.; Birney, E.; Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc. 2009, 4, 1184–1191.

Eynden, J.V.D.; Fierro, A.C.; Verbeke, L.P.C.; Marchal, K. SomInaClust: Detection of cancer genes based on somatic mutation patterns of inactivation and clustering. BMC Bioinform. 2015, 16, 1–12.

Fleischer T., Frigessi A., Johnson K.C., Edvardsen H., Touleimat N., Klajic J., Riis M.L., Haakensen V.D., Wärnberg F., Naume B. et al. Genome-wide DNA methylation profiles in progression to in situ and invasive carcinoma of the breast with impact on gene transcription and prognosis. Genome Biol. 2014; 15:435.

Garon EB, Rizvi NA, Hui R, et al. Pembrolizumab for the treatment of non-small-cell lung cancer [J]. N Engl J Med. 2015;372(21):2018–28.

Gerds, T.A.; Ozenne, B. RiskRegression: Risk Regression Models and Prediction Scores for Survival Analysis with Competing Risks. R Package Version 2020.12.08. 2020. Available online: https://cran.r-project.org/package=riskRegression (accessed on 21 May 2020).

Gentles AJ, Malhotra R, Farshidfar F, Colaprico A, Parker JS, Mose LE, Vo NS, Liu J, Liu Y, Rader J, Dhankani V, Reynolds SM, Bowlby R, Califano A, Cherniack AD, Anastassiou D, Bedognetti D, Mokrab Y, Newman AM, Rao A, Chen K, Krasnitz A, Hu H, Malta TM, Noushmehr H, Pedamallu CS, Bullman S, Ojesina AI, Lamb A, Zhou W, Shen H, Choueiri TK, Weinstein JN, Guinney J, Saltz J, Holt RA, Rabkin CS; Cancer Genome Atlas Research Network, Lazar AJ, Serody JS, Demicco EG, Disis ML, Vincent BG, Shmulevich I. The Immune Landscape of Cancer. Immunity. 2018 Apr 17;48(4):812-830.e14.

Gevaert O., Tibshirani R., Plevritis S.K. Pancancer analysis of DNA methylation-driven genes using MethylMix. Genome Biol. 2015; 16:17.

Goswami S, Chen Y, Anandhan S, Szabo PM, Basu S, Blando JM, Liu W, Zhang J, Natarajan SM, Xiong L, Guan B, Yadav SS, Saci A, Allison JP, Galsky MD, Sharma P. ARID1A mutation plus CXCL13 expression act as combinatorial biomarkers to predict responses to immune checkpoint therapy in mUCC. Sci Transl Med. 2020 Jun 17;12(548):eabc4220.

Grasso CS, Giannakis M, Wells DK, Hamada T, Mu XJ, Quist M, Nowak JA, Nishihara R, Qian ZR, Inamura K, Morikawa T, Nosho K, Abril-Rodriguez G, Connolly C, Escuin-Ordinas H, Geybels MS, Grady WM, Hsu L, Hu-Lieskovan S, Huyghe JR, Kim YJ, Krystofinski P, Leiserson MDM, Montoya DJ, Nadel BB, Pellegrini M, Pritchard CC, Puig-Saus C, Quist EH, Raphael BJ, Salipante SJ, Shin DS, Shinbrot E, Shirts B, Shukla S, Stanford JL, Sun W, Tsoi J, Upfill-Brown A, Wheeler DA, Wu CJ, Yu M, Zaidi SH, Zaretsky JM, Gabriel SB, Lander ES, Garraway LA, Hudson TJ, Fuchs CS, Ribas A, Ogino S, Peters U. Genetic

Mechanisms of Immune Evasion in Colorectal Cancer. Cancer Discov. 2018 Jun;8(6):730-749.

Gu, Z.; Gu, L.; Eils, R.; Schlesner, M.; Brors, B. circlize implements and enhances circular visualization in R. Bioinformatics 2014; 30, 2811–2812.

Gu, Z.; Eils, R.; Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 2016; 32, 2847–2849.

Guo, X., Zhang, Y., Zheng, L. et al. Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. Nat Med 2018; 24, 978–985.

Han, S., Liu, Y., Cai, S.J. *et al.* IDH mutation in glioma: molecular mechanisms and potential therapeutic targets. Br J Cancer 2020; 122, 1580–1589.

Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell 2011;144(5):646–74.

Heagerty, P.J.; Saha-Chaudhuri, P. survivalROC: Time-Dependent ROC Curve Estimation from Censored Survival Data. R Package Version 1.0.3. 2013. Available online: https://cran.r-project.org/package=survivalROC.

Hull, R.M., Cristina, C., Jonathan, V., J.C, H., Mark S.. Environmental change drives accelerated adaptation through stimulated copy number variation. PLoS Biol., 2020; 15, p. e2001333.

Irizarry R.A., Ladd-Acosta C., Wen B., Wu Z., Montano C., Onyango P., Cui H., Gabo K., Rongione M., Webster M. The human colon cancer methylome shows similar hypo-and hypermethylation at conserved tissue-specific CpG island shores. Nat. Genet. 2009; 41:178–186.

Jin, S., Sun, Y., Liang, X. et al. Emerging new therapeutic antibody derivatives for cancer treatment. Sig Transduct Target Ther 7, 2022; 39.

Johnson, D.B., Nebhan, C.A., Moslehi, J.J. et al. Immune-checkpoint inhibitors: long-term implications of toxicity. Nat Rev Clin Oncol 2022; 19, 254–267.

Kassambara, A.; Kosinski, M.; Biecek, P. Survminer: Drawing Survival Curves Using “ggplot2”. R Package Version 0.4.8. 2020. Available online: https://cran.r-project.org/package=survminer (accessed on 21 May 2020).

Kennedy, N. Forestmodel: Forest Plots from Regression Models. R Package Version 0.6.2. 2020. Available online: https://cran.r-project.org/package=forestmodel (accessed on 21 May 2020).

Kristensen V.N., Lingjærde O.C., Russnes H.G., Vollan H. K.M., Frigessi A., Børresen-Dale A.-L.. Principles and methods of integrative genomic analyses in cancer. Nat. Rev. Cancer. 2014; 14:299–313.

Lahiri A, Maji A, Potdar PD, Singh N, Parikh P, Bisht B, Mukherjee A, Paul MK. Lung cancer immunotherapy: progress, pitfalls, and promises. Mol Cancer. 2023 Feb 21;22(1):40.

Lawrence, M.; Huber, W.; Pagès, H.; Aboyoun, P.; Carlson, M.; Gentleman, R.; Morgan, M.T.; Carey, V.J. Software for Computing and Annotating Genomic Ranges. PLoS Comput. Biol. 2013, 9, e1003118.

Leader A. M. et al. "Single-cell analysis of human non-small cell lung cancer lesions refines tumor classification and patient stratification," Cancer Cell, 2021; 39, 12: 1594-1609.e12.

Li, C. Y. Hou, J. Xu, A. Zhang, Z. Liu, F. Qi, Z. Yang, K. Chen, S. Liu, H. Huang. A direct test of selection in cell populations using the diversity in gene expression within tumors. Mol. Biol. Evol. 2017; p. 7

Li, Feng, Caichen Li, Xiuyu Cai, Zhanhong Xie, Liquan Zhou, Bo Cheng, Ran Zhong, Shan Xiong, Jianfu Li, Zhuxing Chen, Ziwen Yu, Jianxing He, Wenhua Liang. The association between CD8+ tumor-infiltrating lymphocytes and the clinical outcome of cancer immunotherapy: A systematic review and meta-analysis. eClinicalMedicine, 2021, 101134.

Liao R, Ma QZ, Zhou CY, Li JJ, Weng NN, Yang Y, Zhu Q. Identification of biomarkers related to Tumor-Infiltrating Lymphocytes (TILs) infiltration with gene co-expression network in colorectal cancer. Bioengineered. 2021 Dec;12(1):1676-1688.

Liu Y, Zhou H, Zheng J, Zeng X, Yu W, Liu W, Huang G, Zhang Y, Fu W. Identification of Immune-Related Prognostic Biomarkers Based on the Tumor Microenvironment in 20 Malignant Tumor Types with Poor Prognosis. Front Oncol. 2020 Jul 31; 10:1008.

Lopez de Rodas M, Nagineni V, Ravi A, Datar IJ, Mino-Kenudson M, Corredor G, Barrera C, Behlman L, Rimm DL, Herbst RS, Madabhushi A, Riess JW, Velcheti V, Hellmann MD, Gainor J, Schalper KA. Role of tumor infiltrating lymphocytes and spatial immune heterogeneity in sensitivity to PD-1 axis blockers in non-small cell lung cancer. J Immunother Cancer. 2022 Jun;10(6):e004440.

Mayakonda, A.; Lin, D.-C.; Assenov, Y.; Plass, C.; Koeffler, H.P. Maftools: Efficient and comprehensive analysis of somatic variants in cancer. Genome Res. 2018, 28, 1747–1756.

McCarthy, D.J.; Chen, Y.; Smyth, G.K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 2012, 40, 4288–4297.

Meyer, D.; Zeileis, A.; Hornik, K. The Strucplot Framework: Visualizing Multi-way Contingency Tables withvcd. J. Stat. Softw. 2006, 17, 1–48.

Meyer, D.; Zeileis, A.; Hornik, K. Vcd: Visualizing Categorical Data. R Package Version 1.4-8. 2020. Available online: https://cran.r-project.org/package=vcd (accessed on 21 May 2020).

Miao YD, Wang JT, Yang Y, Ma XP, Mi DH. Identification of prognosis-associated immune genes and exploration of immune cell infiltration in colorectal cancer. Biomark Med. 2020 Oct;14(14):1353-1369.

Morganella, S.; Pagnotta, S.M.; Ceccarelli, M. GAIA: An R Package for Genomic Analysis of Significant Chromosomal Aberrations. R Package Version 2.32.0. 2020. Available online: https://bioconductor.org/packages/gaia (accessed on 21 May 2020).

Newman, A. M. et al. Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods 2015; 12, 453–457.

Pardoll DM. The blockade of immune checkpoints in cancer immunotherapy. NatRev Cancer 2012;12(4):252–64.

Patil, I. Visualizations with statistical details: The 'ggstatsplot' approach. Journal of Open Source Software, 2021, 6(61), 3167.

Peng J, Zou D, Gong W, Kang S, Han L. Deep neural network classification based on somatic mutations potentially predicts clinical benefit of immune checkpoint blockade in lung adenocarcinoma. Oncoimmunology. 2020 Feb 29;9(1):1734156.

Pinato DJ, Howlett S, Ottaviani D, Urus H, Patel A, Mineo T, Brock C, Power D, Hatcher O, Falconer A, Ingle M, Brown A, Gujral D, Partridge S, Sarwar N, Gonzalez M, Bendle M, Lewanski C, Newsom-Davis T, Allara E, Bower M. Association of Prior Antibiotic Treatment With Survival and Response to Immune Checkpoint Inhibitor Therapy in Patients With Cancer. JAMA Oncol. 2019 Dec 1;5(12):1774-1778.

Pinyol R, Montal R, Bassaganyas L, Sia D, Takayama T, Chau GY, Mazzaferro V, Roayaie S, Lee HC, Kokudo N, Zhang Z, Torrecilla S, Moeini A, Rodriguez-Carunchio L, Gane E, Verslype C, Croitoru AE, Cillo U, de la Mata M, Lupo L, Strasser S, Park JW, Camps J, Solé M, Thung SN, Villanueva A, Pena C, Meinhardt G, Bruix J, Llovet JM. Molecular predictors of prevention of recurrence in HCC with sorafenib as adjuvant treatment and prognostic factors in the phase 3 STORM trial. Gut. 2019 Jun;68(6):1065-1075.

Qian Y, Daza J, Itzel T, Betge J, Zhan T, Marmé F, Teufel A. Prognostic Cancer Gene Expression Signatures: Current Status and Challenges. Cells. 2021 Mar 15;10(3):648.

Qianxing Mo, Sijian Wang, Venkatraman E. Seshan, Adam B. Olshen, Nikolaus Schultz, Chris Sander, R. Scott Powers, Marc Ladanyi, and Ronglai Shen. Pattern discovery and cancer gene identification in integrated cancer genomic data. Proc. Natl. Acad. Sci. USA 2013; 110(11):4245-50.

Qianxing Mo, Ronglai Shen, Cui Guo, Marina Vannucci, Keith S Chan, Susan G Hilsen-beck. A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. Biostatistics 2018; 19(1):71-86.

Ren Q, Zhu P, Zhang H, Ye T, Liu D, Gong Z, Xia X. Identification and validation of stromal-tumor microenvironment-based subtypes tightly associated with PD-1/PD-L1 immunotherapy and outcomes in patients with gastric cancer. Cancer Cell Int. 2020 Mar 24;20:92.

Ren, Y., S. Huang, C. Dai, D. Xie, L. Zheng, H. Xie, H. Zheng, Y. She, F. Zhou, Y. Wang. Germline predisposition and copy number alteration in pre-stage lung adenocarcinomas presenting as ground-glass nodules. Front. Oncol., 2019; 9: 288

Richardson S., Tseng G.C., Sun W. Statistical methods in integrative genomics. Annu. Rev. Stat. Appl. 2016; 3:181–209.

Ritchie, M.E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C.W.; Shi, W.; Smyth, G.K. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015, 43, e47.

Rizvi N. A. et al. Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer. Science, 2015; 348, 124-128.

Robert C, Long GV, Brady B, et al. Nivolumab in previously untreated melanoma without BRAF mutation [J]. N Engl J Med. 2015;372(4):320–30.

Scott, A., Wolchok, J. & Old, L. Antibody therapy of cancer. Nat Rev Cancer 2012; 12, 278–287.

Sengupta, S., S.Q. Sun, K.L. Huang, C. Oh, M.H. Bailey, R. Varghese, M.A. Wyczalkowski, J. Ning, P. Tripathi, J.F. Mcmichael. Integrative omics analyses broaden treatment targets in human cancer. Genome Med., 2018; 10:60.

Shen H., Laird P.W. Interplay between the cancer genome and epigenome. Cell. 2013; 153:38–55.

Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009 Nov 15;25(22):2906-12.

Shin DS, Zaretsky JM, Escuin-Ordinas H, Garcia-Diaz A, Hu-Lieskovan S, Kalbasi A, Grasso CS, Hugo W, Sandoval S, Torrejon DY, Palaskas N, Rodriguez GA, Parisi G, Azhdam A, Chmielowski B, Cherry G, Seja E, Berent-Maoz B, Shintaku IP, Le DT, Pardoll DM, Diaz LA Jr, Tumeh PC, Graeber TG, Lo RS, Comin-Anduix B, Ribas A. Primary Resistance to PD-1 Blockade Mediated by JAK1/2 Mutations. Cancer Discov. 2017 Feb;7(2):188-201.

Silva, T.C.; Colaprico, A.; Olsen, C.; D'Angelo, F.; Bontempi, G.; Ceccarelli, M.; Noushmehr, H. TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages. F1000Research 2016, 5, 1542.

Simon, N.; Friedman, J.H.; Hastie, T.; Tibshirani, R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. J. Stat. Softw. 2011, 39, 1–13.

Sivan A, Corrales L, Hubert N, Williams JB, Aquino-Michaels K, Earley ZM, Benyamin FW, Lei YM, Jabri B, Alegre ML, Chang EB, Gajewski TF. Commensal Bifidobacterium promotes antitumor immunity and facilitates anti-PD-L1 efficacy. Science. 2015 Nov 27;350(6264):1084-9.

Szender JB, Papanicolau-Sengos A, Eng KH, Miliotto AJ, Lugade AA, Gnjatic S, Matsuzaki J, Morrison CD, Odunsi K. NY-ESO-1 expression predicts an aggressive phenotype of ovarian cancer. Gynecol Oncol. 2017 Jun;145(3):420-425.

Tate, J.G.; Bamford, S.; Jubb, H.C.; Sondka, Z.; Beare, D.M.; Bindal, N.; Boutselakis, H.; Cole, C.G.; Creatore, C.; Dawson, E.; et al. COSMIC: The Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. 2018, 47, D941–D947.

Taube JM, Klein A, Brahmer JR, Xu H, Pan X, Kim JH, Chen L, Pardoll DM, Topalian SL, Anders RA. Association of PD-1, PD-1 ligands, and other features of the tumor immune microenvironment with response to anti-PD-1 therapy. Clin Cancer Res. 2014 Oct 1;20(19):5064-74.

TCGARN. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 2012; 489, 519–525.

TCGARN. Integrated genomic characterization of endometrial carcinoma. *Nature* 2013; 497, 67–73.

TCGARN. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* 2014; 513, 202–209.

TCGARN. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* 2014; 511, 543–550.

Therneau, T. A Package for Survival Analysis in R. R Package Version 3.2-7. 2020. Available online: https://cran.r-project.org/package=survival (accessed on 21 May 2020).

Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, Porta-Pardo E, Gao GF, Plaisier CL, Eddy JA, Ziv E, Culhane AC, Paull EO, Sivakumar IKA, Tumeh PC, Harview CL, Yearley JH, Shintaku IP, Taylor EJ, Robert L, Chmielowski B, Spasic M, Henry G, Ciobanu V, West AN, Carmona M, Kivork C, Seja E, Cherry G, Gutierrez AJ, Grogan TR, Mateus C, Tomasic G, Glaspy JA, Emerson RO, Robins H, Pierce RH, Elashoff DA, Robert C, Ribas A. PD-1 blockade induces responses by inhibiting adaptive immune resistance. Nature. 2014 Nov 27;515(7528):568-71.

Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A, Olsson I, Edlund K, Lundberg E, Navani S, Szigyarto CA, Odeberg J, Djureinovic D, Takanen JO, Hober S, Alm T, Edqvist PH, Berling H, Tegel H, Mulder J, Rockberg J, Nilsson P, Schwenk JM, Hamsten M, von Feilitzen K, Forsberg M, Persson L, Johansson F, Zwahlen M, von Heijne G, Nielsen J, Pontén F. Proteomics. Tissue-based map of the human proteome. Science. 2015 Jan 23;347(6220):1260419.

Vadakekolathu J, Lai C, Reeder S, Church SE, Hood T, Lourdusamy A, Rettig MP, Aldoss I, Advani AS, Godwin J, Wieduwilt MJ, Arellano M, Muth J, Yau TO, Ravandi F, Sweet K, Altmann H, Foulds GA, Stölzel F, Middeke JM, Ciciarello M, Curti A, Valk PJM, Löwenberg B, Gojo I, Bornhäuser M, DiPersio JF, Davidson-Moncada JK, Rutella S. TP53 abnormalities correlate with immune infiltration and associate with response to flotetuzumab immunotherapy in AML. Blood Adv. 2020 Oct 27;4(20):5011-5024.

Vrahatis, A.G.; Balomenos, P.; Tsakalidis, A.K.; Bezerianos, A. DEsubs: An R package for flexible identification of differentially expressed subpathways using RNA-seq experiments. Bioinformatics 2016; 32, 3844–3846.

Wiesweg M, Mairinger F, Reis H, Goetz M, Kollmeier J, Misch D, Stephan-Falkenau S, Mairinger T, Walter RFH, Hager T, Metzenmacher M, Eberhardt WEE, Zaun G, Köster J, Stuschke M, Aigner C, Darwiche K, Schmid KW, Rahmann S, Schuler M. Machine learning reveals a PD-L1-independent prediction of response to immunotherapy of non-small cell lung cancer by gene expression context. Eur J Cancer. 2020 Nov;140:76-85

Xie F, Zhang J, Wang J, Reuben A, Xu W, Yi X, Varn FS, Ye Y, Cheng J, Yu M, Wang Y, Liu Y, Xie M, Du P, Ma K, Ma X, Zhou P, Yang S, Chen Y, Wang G, Xia X, Liao Z, Heymach JV, Wistuba II, Futreal PA, Ye K, Cheng C, Xia T. Multifactorial Deep Learning Reveals Pan-Cancer Genomic Tumor Clusters with Distinct Immunogenomic Landscape and Response to Immunotherapy. Clin Cancer Res. 2020 Jun 15;26(12):2908-2920.

Xu, Y. Q. Dong, F. Li, Y. Xu, Y. Zhang. Identifying subpathway signatures for individualized anticancer drug response by integrating multi-omics data. J. Transl. Med. 2019, p. 17

Yang B, Su Z, Chen G, Zeng Z, Tan J, Wu G, Zhu S, Lin L. Identification of prognostic biomarkers associated with metastasis and immune infiltration in osteosarcoma. Oncol Lett. 2021 Mar;21(3):180

Yang, X., L. Gao, S. Zhang. Comparative pan-cancer DNA methylation analysis reveals cancer common and specific patterns. Brief. Bioinformatics, 2020.

Yang, X., Y. Chu, R. Zhang, Y. Han, L. Zhang, Y. Fu, D. Li, R. Peng, D. Li, J. Ding. Technical validation of a next-generation sequencing assay for detecting clinically relevant levels of breast cancer–related single-nucleotide variants and copy number variants using simulated cell-free DNA. J. Mol. Diagn., 2017; 19: 525-536

Yao Y, Yan Z, Lian S, Wei L, Zhou C, Feng D, Zhang Y, Yang J, Li M, Chen Y. Prognostic value of novel immune-related genomic biomarkers identified in head and neck squamous cell carcinoma. J Immunother Cancer. 2020 Jul;8(2): e000444.

Yu, G.; Wang, L.-G.; Han, Y.; He, Q.-Y. clusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters. OMICS J. Integr. Biol. 2012, 16, 284–287.

Yu, G. Enrichplot: Visualization of Functional Enrichment Result. R Package Version 1.8.1. 2020. Available online: https://github.com/GuangchuangYu/enrichplot (accessed on 21 May 2020).

Zaretsky JM, Garcia-Diaz A, Shin DS, Escuin-Ordinas H, Hugo W, Hu-Lieskovan S, Torrejon DY, Abril-Rodriguez G, Sandoval S, Barthly L, Saco J, Homet Moreno B, Mezzadra R, Chmielowski B, Ruchalski K, Shintaku IP, Sanchez PJ, Puig-Saus C, Cherry G, Seja E, Kong X, Pang J, Berent-Maoz B, Comin-Anduix B, Graeber TG, Tumeh PC, Schumacher TN, Lo RS, Ribas A. Mutations Associated with Acquired Resistance to PD-1 Blockade in Melanoma. N Engl J Med. 2016 Sep 1; 375(9):819-29.

Zengin T, Önal-Süzek T. Comprehensive Profiling of Genomic and Transcriptomic Differences between Risk Groups of Lung Adenocarcinoma and Lung Squamous Cell Carcinoma. Journal of Personalized Medicine. 2021; 11(2):154.

Zengin T, Abak Masud B, Önal-Süzek T. TCGAnalyzeR: a web application for integrative visualization of molecular and clinical data of cancer patients for cohort and associated gene discovery. bioRxiv, 2023.

Zhang, J.; Jin, Z. Cutoff: Seek the Significant Cutoff Value. R Package Version 1.3. 2019. Available online: https://cran.r-project.org/package=cutoff (accessed on 21 May 2020).

Zhang, J.; Jin, Z. Ggrisk: Risk Score Plot for Cox Regression. R Package Version 1.2. 2020. Available online: https://cran.r-project.org/package=ggrisk (accessed on 21 May 2020).

Zhao S, Yu M. Identification of MMP1 as a Potential Prognostic Biomarker and Correlating with Immune Infiltrates in Cervical Squamous Cell Carcinoma. DNA Cell Biol. 2020 Feb; 39(2):255-272.

Zhu S, Han X, Qiao X, Chen S. The Immune Landscape and Prognostic Immune Key Genes Potentially Involved in Modulating Synaptic Functions in Prostate Cancer. Front Oncol. 2020 Aug 14; 10:1330.

APPENDIX A.

ARTICLE I

Comprehensive Profiling of Genomic and Transcriptomic Differences between Risk Groups of Lung Adenocarcinoma and Lung Squamous Cell Carcinoma

Talip Zengin and Tuğba Önal-Süzek

Journal of Personalized Medicine, 2021, doi: 10.3390/jpm11020154

Journal of
Personalized
Medicine

Article

Comprehensive Profiling of Genomic and Transcriptomic Differences between Risk Groups of Lung Adenocarcinoma and Lung Squamous Cell Carcinoma

Talip Zengin [1,2] and Tuğba Önal-Süzek [2,3,*]

[1] Department of Molecular Biology and Genetics, Muğla Sıtkı Koçman University, 48000 Muğla, Turkey; talipzengin@mu.edu.tr
[2] Department of Bioinformatics, Muğla Sıtkı Koçman University, 48000 Muğla, Turkey
[3] Department of Computer Engineering, Muğla Sıtkı Koçman University, 48000 Muğla, Turkey
* Correspondence: tugbasuzek@mu.edu.tr

Abstract: Lung cancer is the second most frequently diagnosed cancer type and responsible for the highest number of cancer deaths worldwide. Lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) are subtypes of non-small-cell lung cancer which has the highest frequency of lung cancer cases. We aimed to analyze genomic and transcriptomic variations including simple nucleotide variations (SNVs), copy number variations (CNVs) and differential expressed genes (DEGs) in order to find key genes and pathways for diagnostic and prognostic prediction for lung adenocarcinoma and lung squamous cell carcinoma. We performed a univariate Cox model and then lasso-regularized Cox model with leave-one-out cross-validation using The Cancer Genome Atlas (TCGA) gene expression data in tumor samples. We generated 35- and 33-gene signatures for prognostic risk prediction based on the overall survival time of the patients with LUAD and LUSC, respectively. When we clustered patients into high- and low-risk groups, the survival analysis showed highly significant results with high prediction power for both training and test datasets. Then, we characterized the differences including significant SNVs, CNVs, DEGs, active subnetworks, and the pathways. We described the results for the risk groups and cancer subtypes separately to identify specific genomic alterations between both high-risk groups and cancer subtypes. Both LUAD and LUSC high-risk groups have more downregulated immune pathways and upregulated metabolic pathways. On the other hand, low-risk groups have both up- and downregulated genes on cancer-related pathways. Both LUAD and LUSC have important gene alterations such as CDKN2A and CDKN2B deletions with different frequencies. SOX2 amplification occurs in LUSC and PSMD4 amplification in LUAD. EGFR and KRAS mutations are mutually exclusive in LUAD samples. EGFR, MGA, SMARCA4, ATM, RBM10, and KDM5C genes are mutated only in LUAD but not in LUSC. CDKN2A, PTEN, and HRAS genes are mutated only in LUSC samples. The low-risk groups of both LUAD and LUSC tend to have a higher number of SNVs, CNVs, and DEGs. The signature genes and altered genes have the potential to be used as diagnostic and prognostic biomarkers for personalized oncology.

Keywords: TCGA; non-small-cell lung cancer; lung adenocarcinoma (LUAD); lung squamous cell carcinoma (LUSC); differential expression; SNV; CNV; risk group; signature; survival

check for updates

Citation: Zengin, T.; Önal-Süzek, T. Comprehensive Profiling of Genomic and Transcriptomic Differences between Risk Groups of Lung Adenocarcinoma and Lung Squamous Cell Carcinoma. *J. Pers. Med.* **2021**, *11*, 154. https://doi.org/10.3390/jpm11020154

Academic Editor: Raghu Sinha

Received: 30 December 2020
Accepted: 19 February 2021
Published: 23 February 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Lung cancer is the second most frequently diagnosed cancer type and the leading cause of cancer-related mortality worldwide [1]. Lung cancer treatments used in the clinic are surgery, radiotherapy, chemotherapy, targeted therapy, and emerging immunotherapy. The clinical treatment decisions are made based on tumor stage, histology, genetic alterations of a few driver oncogenes for targeted therapies, and patient's condition [2]. However, most of the patients are diagnosed at an advanced and metastatic stage, with high mortality and poor benefit from therapies [3]. Although the targeted therapeutics and immunotherapeutics including immune-checkpoint inhibitors are introduced for patients at an advanced stage, these options are beneficial only for limited subsets of patients and these patients still can develop resistance [4]. Therefore, the majority of patients with advanced-stage lung cancer die within 5 years of diagnosis [5].

Histologically there are four major types of lung cancer, including small-cell carcinoma (SCLC), and adenocarcinoma, squamous cell carcinoma, large cell carcinoma as grouped non-small-cell carcinoma (NSCLC). Lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) account for 50% and 23% of all lung cancers, respectively [6]. Lung cancer is both histologically and molecularly heterogeneous disease and characterizing the genomics and transcriptomics of its nature is very important for effective therapies. Lung cancer has many subtypes with distinct genetic characteristics, resulting in intra-tumoral heterogeneity [7].

The Cancer Genome Atlas (TCGA) database serves different types of data such as transcriptome profiling, simple nucleotide variation, copy number variation, DNA methylation, clinical and biospecimen data of 84,392 cancer patients with 68 primary sites [8]. The Cancer Genome Atlas Research Network reported molecular profiling of 230 lung adenocarcinoma samples using mRNA, microRNA and DNA sequencing integrated with copy number, methylation and proteomic analyses. They identified 18 significantly mutated genes, including TP53, KRAS which is mutually exclusive with EGFR, BRAF, PIK3CA, MET, STK11, KEAP1, NF1, RB1, CDKN2A, GTPase gene RIT1, including activating mutations and MGA including loss-of-function mutations. DNA and mRNA sequence from the same tumor highlighted splicing alterations including exon 14 skipping in MET mRNA in 4% of cases. They also showed DNA hyper-methylation of several key genes: CDKN2A, GATA2, GATA4, GATA5, HIC1, HOXA9, HOXD13, RASSF1, SFRP1, SOX17, WIF1, and MYC over-expression was significantly associated with the hyper-methylation phenotype as well [9].

The Cancer Genome Atlas Research Network also profiled 178 lung squamous cell carcinomas and detected mutations in 11 genes, including mutations in TP53 (81%), CDKN2A, PTEN, PIK3CA, KEAP1, MLL2, HLA-A, NFE2L2, RB1, NOTCH1 including truncating mutations and loss-of-function mutations in the HLA-A class I major histocompatibility gene. They identified altered pathways such as NFE2L2 and KEAP1 in 34%, squamous differentiation genes in 44%, PI3K pathway genes in 47%, and CDKN2A and RB1 in 72% of tumors. CNV analysis revealed the amplification of NFE2L2, MYC, CDK6, MDM2, BCL2L1 and EYS, and deletions of FOXP1, PTEN and NF1 genes with previously identified CNV genes, SOX2, PDGFRA, KIT, EGFR, FGFR1, WHSC1L1, CCND1, and CDKN2A. They identified overexpression and amplification of SOX2 and TP63, loss-of-function mutations in NOTCH1, NOTCH2 and ASCL4 and focal deletions in FOXP1 which have known roles in squamous cell differentiation. CDKN2A is downregulated in over 70% of samples through epigenetic silencing by methylation (21%), inactivating mutation (18%), exon 1β skipping (4%), or homozygous deletion (29%) [10].

Recently, many studies have been published on gene expression signatures predicting the survival risk of patients with lung adenocarcinoma. These recent studies have been mostly using TCGA data, but their methods generated different gene signatures. Seven-gene expression signature including ASPM, KIF15, NCAPG, FGFR1OP, RAD51AP1, DLGAP5 and ADAM10 genes, was obtained for early stage cases from seven published lung adenocarcinoma cohorts and the signature showed high hazard rations in Cox regression analysis [11]. Shukla et al. developed TCGA RNAseq data-based prognostic signature including four protein-coding genes RHOV, CD109, FRRS1, and the lncRNA gene LINC00941, which showed high hazard ratios for stage I, EGFR wild-type, and EGFR mutant groups [12]. A prognostic signature that was independent of other clinical factors, was developed and validated based on the TCGA data. Patients were grouped into risk groups using signature genes, and patients with high-risk scores tended to have poor survival rate at 1-, 3- and 5-year follow-up. The developed eight-gene signature including

TTK, HMMR, ASPM, CDCA8, KIF2C, CCNA2, CCNB2, and MKI67 were highly expressed in A549 and PC-9 cells [13].

Twelve-gene signature (RPL22, VEGFA, G0S2, NES, TNFRSF25, DKFZP586P0123, COL8A2, ZNF3, RIPK5, RNFT2, ARHGEF12 and PTPN20A/B) was established by using published microarray dataset from 129 patients and the signature was independently prognostic for lung squamous carcinoma but not for lung adenocarcinoma [14]. A four-gene clustering model in 14-Genes (DPPA, TTTY16, TRIM58, HKDC1, ZNF589, ALDH7A1, LINC01426, IL19, LOC101928358, TMEM92, HRASLS, JPH1, LOC100288778, GCGR) was established and these genes plays role in positive regulation of ERK1 and ERK2 cascade, angiogenesis, platelet degranulation, cell–matrix adhesion, extracellular matrix organization and macrophage activation [15].

Lu et.al. identified differentially expressed genes between lung adenocarcinoma and lung squamous cell carcinoma by using microarray data from the Gene Expression Omnibus database. They identified 95 upregulated and 241 downregulated DEGs in lung adenocarcinoma samples, and 204 upregulated and 285 downregulated DEGs in lung squamous cell carcinoma samples, compared to the normal lung tissue samples. The genes play role in cell-cycle, DNA replication and mismatch repair. The top five genes from global network, HSP90AA1, BCL2, CDK2, KIT and HDAC2 have differential expression profiles between lung adenocarcinoma and lung squamous cell carcinoma [16]. Recently, Wu et.al. identified diagnostic and prognostic genes for lung adenocarcinoma and squamous cell carcinoma by using weighted gene expression profiles. The five-gene diagnostic signature including KRT5, MUC1, TREM1, C3 and TMPRSS2 and the five-gene prognostic signature including ADH1C, AZGP1, CLU, CDK1 and PEG10 obtained a log-rank P-value of 0.03 and a C-index of 0.622 on the test set [17].

A considerable number of genetic and transcriptomic alterations have been identified in mostly LUAD and poorly in LUSC. Although many gene expression signatures have been identified in LUAD recently, there is less work on LUSC expression signatures. Additionally, the molecular differences between risk groups of LUAD and LUSC have not yet been systematically described. In this study, we aimed to identify the genomic and transcriptomic differences between risk groups of lung adenocarcinoma and lung squamous cell carcinoma. We performed a univariate Cox model and then Lasso-Regularized Cox Model with Leave-One-Out Cross-Validation (LOOCV) by using TCGA gene expression data in tumor samples, and identified best gene signatures to cluster patients into low- and high-risk groups. We generated 35- and 33-gene signatures for prognostic risk prediction based on the overall survival time of the patients with LUAD and LUSC. When we clustered patients into high- and low-risk groups, the survival analysis showed highly significant results for both training and test datasets. Then, we characterized the differences including significant SNVs, CNVs, DEGs and active subnetwork DEGs between risk groups in LUAD and LUSC.

2. Materials and Methods

2.1. Data

Simple Nucleotide Variation (SNV), Transcriptome Profiling, Copy Number Variation (CNV) and Clinical data of patients who have all of these data types in LUAD and LUSC projects, was downloaded separately using *TCGAbiolinks* R package [18]. Using the same package and the reference of hg38; Simple Nucleotide Variations (SNVs) and Copy Number Variations (CNVs); and transcriptomic variations were processed to identify the genomic alterations of the LUAD and LUSC patients (Table 1). The method described below can be found as flowchart in Figure S1.

Table 1. Summary of clinical variables of train and test group of patients with LUAD and LUSC analyzed in the study.

	LUAD		**LUSC**	
Category	**Train Group (n: 436)**	**Test Group (n: 56)**	**Train Group (n: 431)**	**Test Group (n: 47)**
Age at diagnosis (median; range)	66; 33–88	66.5; 42–86	68; 39–90	69; 45–85
Gender				
Female	232	33	112	14
Male	204	23	319	33
Tumor stage				
I	241	28	211	25
II	106	13	138	16
III	68	13	76	5
IV	23	2	6	1
Vital status				
Alive	284	30	275	18
Dead	152	26	156	29
Smoked years (median; range)	33; 2–61	31.5; 4–64	40; 8–62	40; 10–60
Smoked packs per year (median; range)	40; 0.15–154	48; 5–94.5	50; 1–240	50; 2–157.5

2.2. Gene Expression Signature Analysis

Clinical data and Gene Expression Quantification data (HTSeq counts) of patients with unpaired RNAseq data (tumor samples without normal samples) was downloaded from the TCGA database using the *TCGAbiolinks* R package. Raw HTSeq counts of tumor samples were normalized by TMM (trimmed mean of M values) method and Log_2 transformed after filtering to remove genes that consistently have zero or low counts. Univariate Cox Proportional Hazards Regression analysis was performed using *survival* R package [19] to identify survival-related genes. For these survival-related potential biomarker genes ($p \leq 0.05$), Lasso-Regularized Cox Model (by using minimum lambda calculated in the model) with Leave-One-Out Cross-Validation (LOOCV) was performed to determine a gene expression signature using *glmnet* R package [20]. Multivariate Cox Regression for the signature genes was performed and the predictive performance of the model was scored using *riskRegression* R package [21]. The risk score of each patient was predicted based on multivariate Cox regression model using the *survival* R package. Patients were clustered into high-risk and the low-risk group based on the best cutoff value for ROC, calculated by *cutoff* R package [22].

For the validation of the gene signature, HTSeq counts belonging to the tumor samples of patients who have paired RNAseq data (tumor samples with the paired adjacent normal samples) were downloaded from the TCGA database, filtered, normalized by TMM method and Log_2 transformed. Multivariate Cox Regression for the signature genes was performed and the predictive performance of the model was scored. The risk score of every patient in the validation group was predicted based on multivariate Cox regression model and each patient was assigned to the high- or low-risk group using the best cutoff value for ROC. These analyses were performed for LUAD and LUSC patients separately.

2.3. Differential Expression Analysis

Gene Expression Quantification data (HTSeq counts) of both the primary tumor (TP) and the paired normal tissue adjacent to the tumor (NT) was downloaded from the TCGA database. Raw HTSeq counts of both tumor and normal samples were normalized by TMM method after filtering to remove genes which have zero or low counts. Differentially expressed ($q < 0.01$) genes were determined using *limma* [23] and *edgeR* [24] R packages by limma-voom method with duplicate-correlation function. HUGO symbols and NCBI Gene identifiers of the differentially expressed genes were downloaded using the *biomaRt* R package. This analysis was performed for high- and low-risk group patients of LUAD and LUSC, separately.

2.4. Active Subnetwork Analysis

Active subnetworks of the differentially expressed genes were determined using *DEsubs* R package [25]. *DEsubs* package accepts the differentially expressed genes output of the *limma* package along with their FDR adjusted p values (q values). *DEsubs* package both computes and plots the active subnetworks. All the plots and computations were generated for the high- and low-risk group patients of the LUAD and LUSC projects, separately.

2.5. Copy Number Variation Analysis

The Copy Number Variation data of the primary tumor samples of patients was downloaded using *TCGAbiolinks* package (Masked Copy Number Segment as data type). The chromosomal regions which are significantly aberrant in tumor samples were determined and plotted by *gaia* R package [26]. Gene enrichment from genomic regions which have significant differential copy number was performed using *GenomicRanges* [27] and *biomaRt* R packages. R codes used in this analysis were modified from the codes presented at "TCGA Workflow" article [28]. All the computations and the plots were generated for the high- and low-risk groups of LUAD and LUSC projects, separately.

2.6. Simple Nucleotide Variations Analysis

The masked Mutation Annotation Format (maf) files of the TCGA mutect2 pipeline in tumor samples were downloaded to obtain the somatic mutations. The maf files are filtered using the *maftools* [29] to obtain the subset of the mutations corresponding to the patient barcodes. Summary plot and oncoplot were generated to summarize the mutation data using *maftools* R package. Somatic mutations were filtered and assigned to either oncogene (OG) or tumor suppressor gene (TSG) groups along with a significance score ($q < 0.05$) using the *SomInaClust* R package [30]. *SomInaClust* computes a background mutation value to identify the hot spots using the known set of somatic mutations in "COSMIC" and the "Cancer Gene Census" (v92) datasets of COSMIC database for GRCh38 [31]. SNV analysis was performed for high- and low-risk group patients of LUAD and LUSC projects, separately.

2.7. Visualization

Scatter plots showing risk score and survival time of patients were generated by *ggrisk* R package [32] and Kaplan–Meier (KM) survival curves were plotted by *survminer* R package [33] displaying the overall survival difference between the risk groups stratified on the proposed gene signature. ROC curves were plotted for the risk scores based on each gene signature using *survivalROC* R package [34]. Univariate and multivariate Cox regression analyses were performed and forest plots were generated for risk score with clinical variables using *survival* and *forestmodel* [35] R packages.

Gene and pathway enrichment analyses were performed by *biomaRt* [36] and *clusterProfiler* [37] R packages and plotted by *enrichplot* R package [38]. Heatmap plots were generated using *ComplexHeatmap* R package [39]. Mosaic plots to compare the categorical variables were generated using the *vcd* R package [40,41].

OncoPrint showing CNVs among patient samples was generated using *ComplexHeatmap* R package. OncoPlot for significant mutated genes was drawn using *maftools*, and oncoPrint showing SNVs and CNVs together was generated using *ComplexHeatmap* R package. Circos plot showing all non-synonymous SNVs in original data of risk groups and significant CNVs at genome-scale were generated using *circlize* R package [42].

All possible relations between DEGs; active subnetwork DEGs; CNV genes; SNV genes of LUAD and LUSC risk groups were identified by using *VennDiagram* R package [43].

3. Results

3.1. Gene Expression Signature Analysis of LUAD and LUSC Patients

In order to identify gene expression prognosis risk model, clinical data and gene expression quantification data of tumor samples of patients with LUAD and h LUSC with unpaired RNAseq data as two separate training groups (Table 1) were downloaded from the TCGA database. A 35-gene expression signature for LUAD and a 33-gene expression signature for LUSC were identified by Lasso-Regularized Cox Model with LOOCV after univariate Cox regression analysis. The risk scores of each patient in training groups and test groups were predicted using signature genes, then patients were clustered into high- and low-risk groups based on the cutoff values.

The genes of the LUAD expression signature model identified are AC005077.4, AC113404.3, ADAMTS15, AL365181.2, ANGPTL4, ASB2, ASCL2, CCDC181, CCL20, CD200R1, CPXM2, DKK1, ENPP5, EPHX1, GNPNAT1, GRIK2, IRX2, LDHA, LDLRAD3, LINC00539, LINC00578, MS4A1, OGFRP1, RAB9B, RGS20, RHOQ, SAMD13, SLC52A1, STAP1, TLE1, U91328.1, WBP2NL, ZNF571-AS1, ZNF682, ZNF835. Twenty-seven of them are protein-coding genes while two of them are long intergenic non-protein coding RNA (LINC00539, LINC00578), one is antisense RNA (ZNF571-AS1), three of them are pseudogenes (AC005077.4, AC113404.3, OGFRP1) and two of them are novel transcripts (AL365181.2, U91328.1) (Table S1). Pathway enrichment analysis by using *clusterProfiler* R package did not give any results for this 35-gene list; therefore, enrichment analysis was performed manually using the online KEGG Mapper tool. The genes play role in metabolic pathways, cancer and immune system-related pathways such as Central carbon metabolism in cancer, Glycolysis, Cholesterol metabolism, Amino sugar and Nucleotide sugar metabolism, HIF-1 signaling pathway, TNF signaling pathway, IL-17 signaling pathway, Chemokine signaling pathway and Wnt signaling pathway (Table S2). Multivariate Cox regression analysis was performed for the signature genes and the predictive performance of the model was scored. The AUC was 0.963 ($p = 1.1 \times 10^{-15}$) for LUAD training group. The risk score of each patient was predicted and patients were clustered into high- and low-risk groups based on the cutoff value. Low- and high-risk groups have different expression patterns of the signature genes and significantly different survival probabilities ($p < 0.0001$). The prediction power of the risk score is around 0.78 (AUC) for 1, 3, 5 and 8 years for LUAD training group (Figure S2). Risk group clustering is independent from tumor stages because risk groups have also significantly different survival probability for each tumor stage (Figure S3). Vital status is highly correlated with risk groups that high-risk group is positively correlated with death ($p = 1.5 \times 10^{-13}$), while only tumor stage IA and III are associated with risk groups (Figure S4). The risk score has highly significant prognostic ability (HR:2.59, $p < 0.001$) when multivariate Cox regression analysis was performed with other clinical variables (Figures S5 and S6).

In order to validate the gene expression signature, gene expression quantification data of tumor samples of patients with LUAD who have paired RNAseq data were downloaded from the TCGA database. The risk scores of each patient in test group were predicted using the gene signature lists and patients were clustered into high- and low-risk groups based on the best cutoff values for ROC. Risk groups have differential signature gene expression patterns; high-risk group has lower survival time and higher number of deaths resulting a significantly different survival probability ($p < 0.0001$). The risk score has high prediction powers, 0.97, 0.92, 0.93 and 0.92 (AUC) for 1, 3, 5 and 8 years, respectively, for LUAD test group (Figure 1).

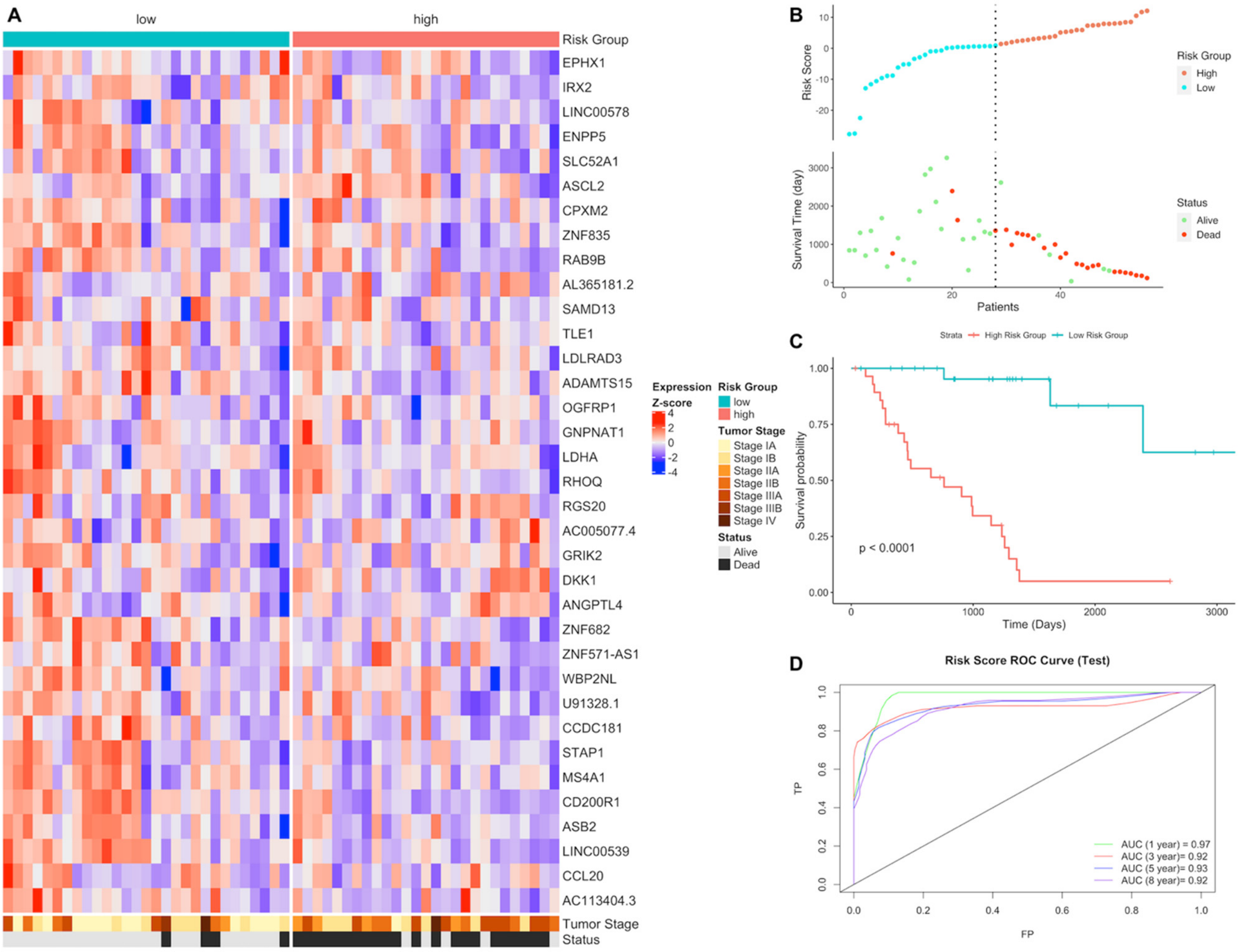


Figure 1. Gene expression signature and risk clustering of LUAD test dataset. Test dataset patients were clustered into high- and low-risk groups based on risk scores of patients calculated by predicting the effect of the signature genes of the signature genes expression on overall survival. (**A**) Expression heatmap of the signature genes in tumor samples of LUAD patients in the test dataset. (**B**) Scatter plot showing risk scores, survival time and separation point of the patients into risk groups. (**C**) KM survival plot showing the overall survival probability between risk groups. (**D**) ROC curve showing prediction power of risk score in the test dataset for 1, 3, 5 and 8 years.

Risk groups have significantly different survival probability for each tumor stage in LUAD test group as well (Figure S7). Vital status is highly correlated with risk groups. The high-risk group is positively correlated with death ($p = 3.87 \times 10^{-7}$), while only tumor stage I is positively associated with low-risk group ($p = 0.016$) (Figure S8). The risk score has highly significant prognostic ability (HR:2.79, $p < 0.001$) as the result of multivariate Cox regression analysis was performed with other clinical variables (Figure S9).

Expression signature model identified for LUSC includes these genes: AC078883.1, AC096677.1, AC106786.1, ADAMTS17, ALDH7A1, ALK, COL28A1, EDN1, FABP6, HKDC1, IGSF1, ITIH3, JHY, KBTBD11, LINC01426, LINC01748, LPAL2, NOS1, PLAAT1, PNMA8B, RGMA, RPL37P6, S100A5, SLC9A9, SNX32, SRP14-AS1, STK24, UBB, UGGT2, WASH8P, Y_RNA, ZNF160, ZNF703. Twenty-three of them are protein coding genes while two of them are long intergenic non-protein coding RNA (LINC01748, LINC01426), one is antisense RNA (SRP14-AS1), three of them are pseudo-genes (LPAL2, RPL37P6, WASH8P), three of them are novel transcripts (AC106786.1, AC096677.1, AC078883.1) and one is Y RNA (Table S3). They play role in mostly in metabolic pathways, cancer and immunity related pathways such as Arginine and proline metabolism, Glycolysis/Gluconeogenesis, HIF-1 signaling pathway, Non-small-cell lung cancer, PD-L1 expression and PD-1 checkpoint pathway in cancer and TGF-beta signaling pathway (Table S4).

The predictive performance score of the signature model is 80.8 (AUC) ($p = 1.3 \times 10^{-6}$) in multivariate Cox regression analysis for LUSC training group. The risk score of each patient was predicted and patients were clustered into high- and low-risk groups based on the cutoff value. Low- and high-risk groups have different expression patterns of the signature genes and significant difference of survival probability ($p < 0.0001$). The AUC values showing prediction power of the risk score are 0.76, 0.82, 0.87 and 0.92 for 1, 3, 5 and 8 years, respectively, for LUSC training group (Figure S10). Risk groups have also significantly different survival probability for tumor stages I, II and III (Figure S11). Risk groups are highly correlated with vital status. The high-risk group has highly significant positive correlation with death ($p = 8.5 \times 10^{-15}$), while low-risk group is negatively correlated. Tumor stages did not show any association with risk groups (Figure S12). The risk score has highly significant prognostic ability (HR:2.85, $p < 0.001$) when multivariate Cox regression analysis was performed with other clinical variables (Figure S13).

In order to validate the gene expression signature for LUSC, gene expression quantification data of tumor samples of patients with LUSC who have paired RNAseq data were downloaded. The risk scores of each patient in LUSC test group were predicted using gene signature lists and patients were clustered into high- and low-risk groups based on the best cutoff values for ROC. Risk groups have differential signature gene expression pattern; high-risk group has lower survival time and higher number of deaths. Risk groups have significantly different survival probability ($p < 0.0001$). The risk score has high prediction powers, 0.93, 0.95, 0.96 and 0.97 (AUC) for 1, 3, 5 and 8 years, respectively, for LUSC test group (Figure 2).

Risk groups have also significantly different survival probability for tumor stages in test group (Figure S14). Vital status is not correlated with risk groups of LUSC test group that number of deaths is higher for high-risk group insignificantly ($p = 0.07$). Tumor stages are not associated with risk groups (Figure S15). The risk score has highly significant prognostic ability (HR:2.66, $p < 0.001$) while other clinical variables have no effect on overall survival in multivariate Cox regression analysis (Figure S16).

The expression gene signatures of LUAD and LUSC do not have any common gene, however they share eight common pathways which are mostly metabolic pathways: Central carbon metabolism in cancer, Glycolysis/Gluconeogenesis, HIF-1 signaling pathway, Pyruvate metabolism, PPAR signaling pathway, Amino sugar and nucleotide sugar metabolism, TNF signaling pathway and Pathways of neurodegeneration—multiple diseases.

3.2. *Differential Expression and Active Subnetwork Analysis of Risk Groups*

Gene expression quantification data of both primary tumor and adjacent normal tissues of patients who have paired RNAseq data (test groups) in LUAD and LUSC projects were downloaded from the TCGA database. Differentially expressed ($q < 0.01$) genes (DEGs) were determined in tumor samples according to normal samples for high- and low-risk patient groups in test sets of LUAD and LUSC, separately. Then, active subnetworks of DEGs in tumor samples were determined using the DEGs with their q values.

In tumor samples of the LUAD low-risk group, the number of the genes which are dysregulated significantly ($q < 0.01$) more than 2-fold is 3615 (2439 down-, 1176 upregulated) while 3610 genes (2239 down-, 1371 upregulated) are dysregulated for the LUAD high-risk group. LUAD low- and high-risk groups have 2745 common differentially expressed genes (Figure S17). The top 20 significant DEGs highlighted as purple at volcano plot in Figure 3A,B are different between LUAD risk groups as dysregulation pattern is different between risk groups albeit the shared 2745 DEGs.

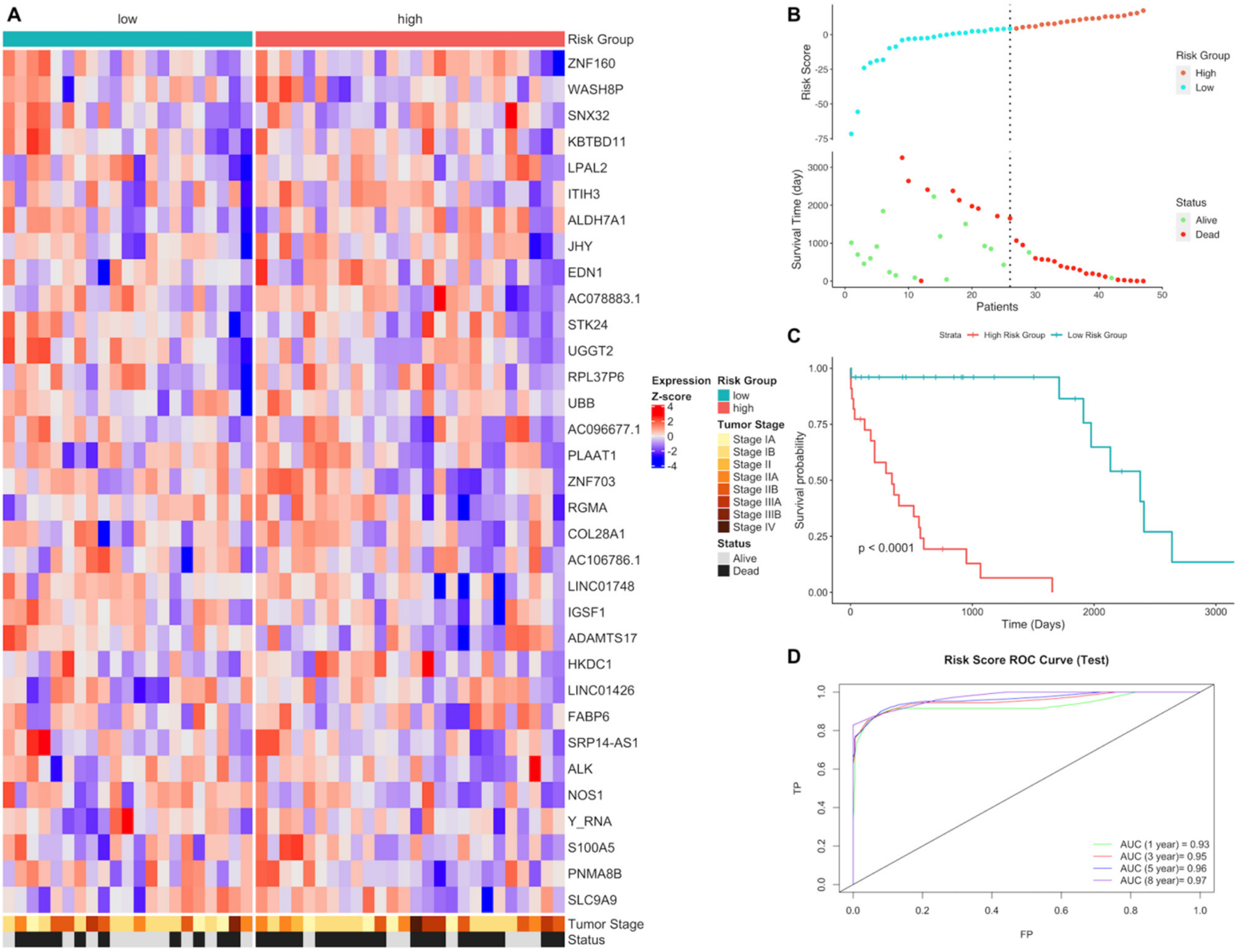


Figure 2. Gene expression signature and risk clustering of LUSC test dataset. Test dataset patients were clustered into high- and low-risk groups based on risk scores of patients calculated by predicting the effect of the signature genes' expression on overall survival. (**A**) Expression heatmap of the signature genes in tumor samples of LUSC patients in the test dataset. (**B**) Scatter plot showing risk scores, survival time and separation point of the patients into risk groups. (**C**) KM survival plot showing the overall survival probability between risk groups. (**D**) ROC curve showing prediction power of risk score in the test dataset for 1, 3, 5, and 8 years.

Seven of the signature genes (GNPNAT1, CCDC181, LDHA, ADAMTS15, IRX2, LINC00578, AC005077.4) are dysregulated in both risk groups. ANGPTL4 is upregulated in the high-risk group while MS4A1, GRIK2, and OGFRP1 are upregulated in the low-risk group.

Risk groups of LUAD share dysregulated pathways (Figure 3C,D), highly related to cancer, such as Cell cycle, Biosynthesis of amino acids and Protein digestion and absorption which are upregulated for both risk groups (Figure S18), on the other hand, they also share ECM–receptor interaction, Cell adhesion molecules pathways with immune system-related pathways such as Complement and coagulation cascades and Cytokine-cytokine receptor interaction which are downregulated for both risk groups (Figure S18). However, the high-risk group has more dysregulated immune system-related pathways such as Allograft rejection, Graft-versus-host disease, Inflammatory bowel disease, Intestinal immune network for IgA production, Rheumatoid arthritis, Staphylococcus aureus infection (Figure 3C,D), which are downregulated pathways in LUAD high-risk group (Figure S18).

Active subnetworks of differentially expressed genes in tumor samples of the LUAD risk groups were identified and low-risk group has 191 genes while high-risk group has 168 genes including 112 common genes, which are acting on active subnetworks (Figure S17).

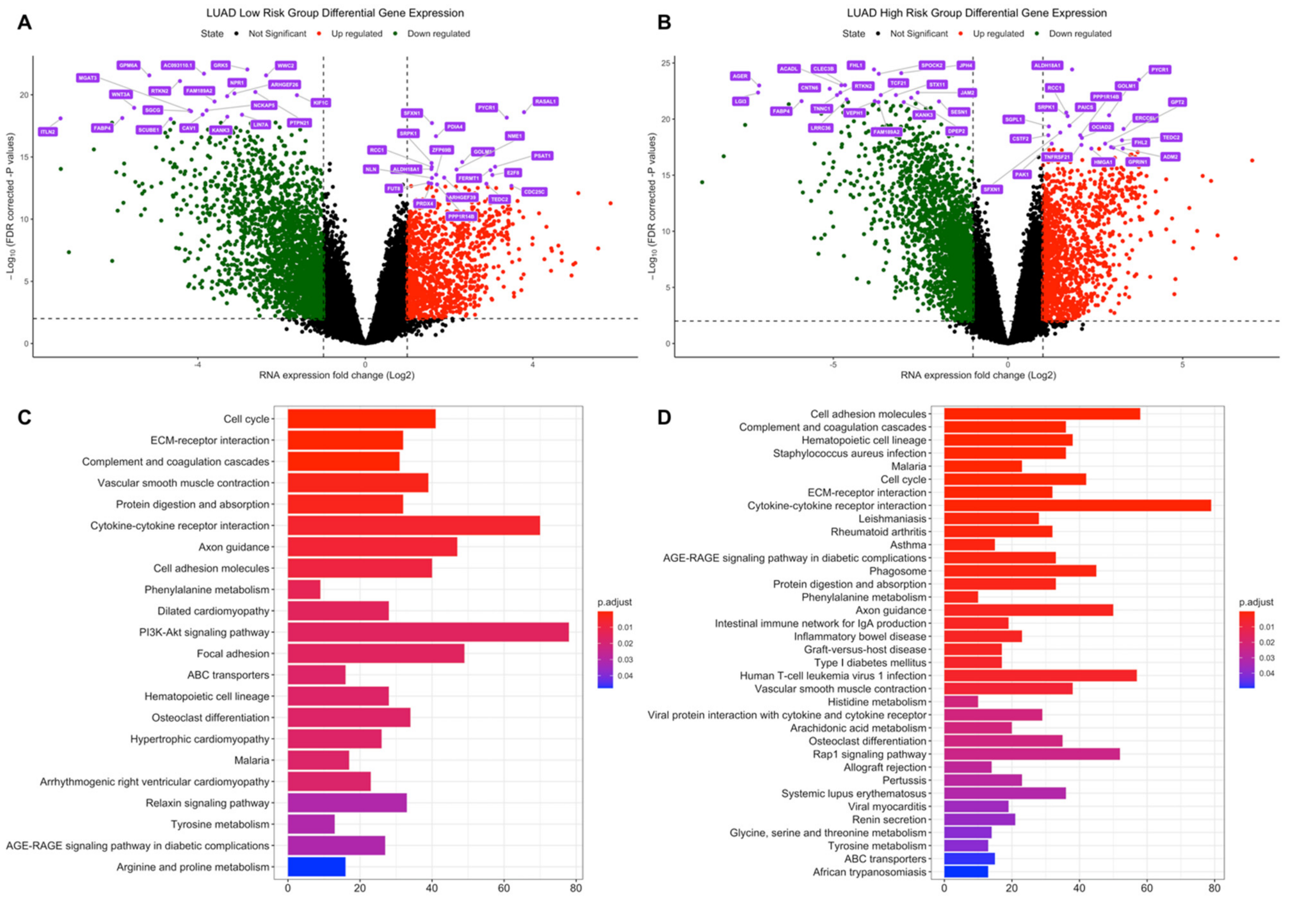


Figure 3. Differential expression analysis of the LUAD risk groups. LUAD test dataset patients were clustered into high- and low-risk groups based on risk scores of patients and differentially expressed genes in tumor samples were determined based on expressions in normal tissues. (**A**) Volcano plot showing differentially expressed genes more than 2-fold (Log_2 =1) for LUAD low-risk group. The top 20 significant downregulated and upregulated genes are highlighted as purple. FDR corrected p-values threshold is 0.01 ($-Log_{10}$ = 2). Red: Upregulated, Green: Downregulated, Black: Not significant or low than 2-fold. (**B**) Volcano plot showing differentially expressed genes more than two-fold (Log_2 = 1) for the LUAD high-risk group. The top 20 significant downregulated and upregulated genes are highlighted as purple. FDR corrected *p*-values threshold is 0.01 ($-Log_{10}$ = 2). Red: Upregulated, Green: Downregulated, Black: Not significant or low than 2-fold. (**C**) Dysregulated pathways of differentially expressed genes for LUAD low-risk group. (**D**) Dysregulated pathways of differentially expressed genes for LUAD high-risk group.

Pathway enrichment of DEGs at active subnetworks shows that the genes playing role in active subnetworks are much more related to cancer pathways such as PI3K-Akt signaling pathway, Ras signaling pathway, Small-cell lung cancer, Breast cancer, Gastric cancer, Proteoglycans in cancer and Rap1 signaling pathway (Figure 4). LUAD risk groups have mostly similar cancer-related active pathways, however only low-risk group has FoxO signaling pathway and TNF signaling pathway while high-risk group has Estrogen signaling pathway, Growth hormone synthesis, secretion, and action with immune system pathways such as Antigen processing and presentation, Intestinal immune network for IgA production and Leukocyte trans-endothelial migration.

The number of dysregulated genes expressed significantly ($q < 0.01$) more than 2-fold in tumor samples of the LUSC low-risk group is 5596 (3394 downregulated, 2202 upregulated) while 5403 genes (3338 downregulated, 2065 upregulated) are dysregulated for LUSC high-risk group. LUSC low- and high-risk groups have 4562 common differentially expressed genes (Figure S17). The top 20 significant DEGs highlighted at volcano plot in Figure 5A,B include common genes and dysregulation pattern is similar between risk groups.

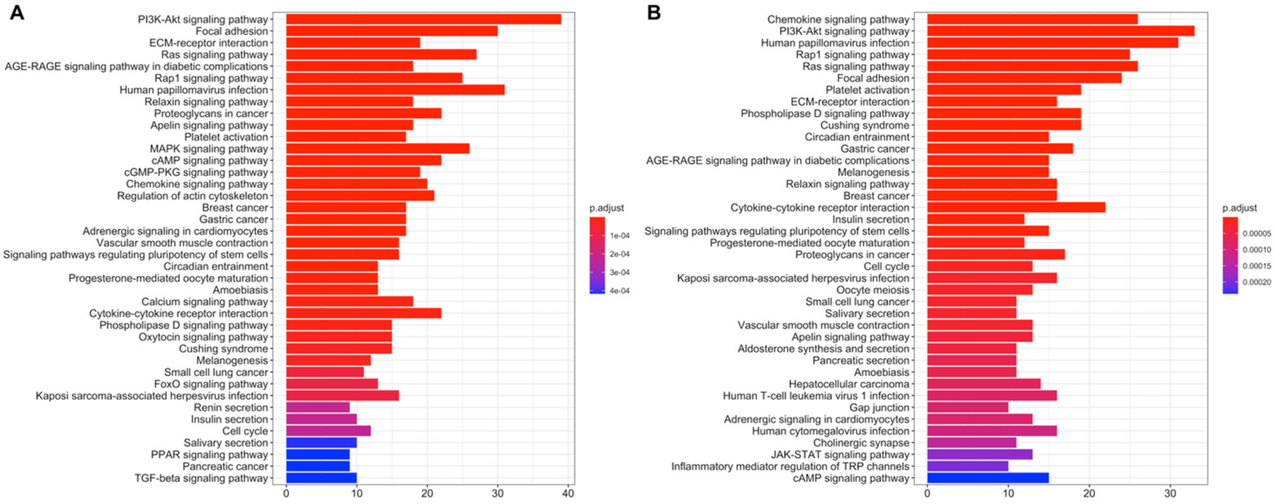


Figure 4. Pathway enrichment of differentially expressed genes at active subnetworks of the LUAD risk groups. Active subnetworks were determined by using differential expression analysis results and pathway enrichment analysis was performed for the genes at subnetworks. (**A**) Pathways of differentially expressed genes in active subnetworks for LUAD low-risk group. (**B**) Pathways of differentially expressed genes in active subnetworks for LUAD high-risk group.

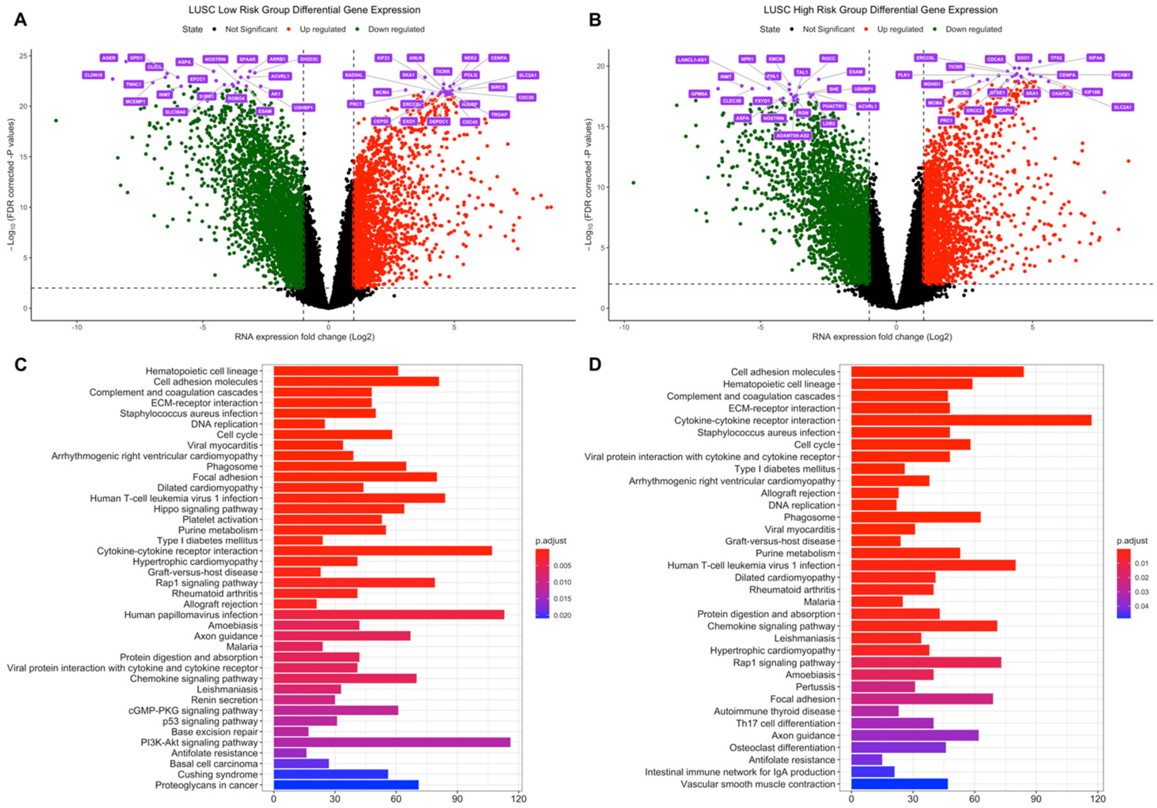


Figure 5. Differential expression analysis of the LUSC risk groups. LUSC test dataset patients were clustered into high- and low-risk groups based on risk scores of patients and differentially expressed genes in tumor samples were determined based on expressions in normal tissues. (**A**) Volcano plot showing differentially expressed genes more than 2-fold ($Log_2 = 1$) for LUSC low-risk group. The top 20 significant downregulated and upregulated genes are highlighted as purple. FDR corrected p-values threshold is 0.01 ($-Log_{10} = 2$). Red: Upregulated, Green: Downregulated, Black: Not significant or low than 2-fold. (**B**) Volcano plot showing differentially expressed genes more than two-fold ($Log_2 = 1$) for LUSC high-risk group. The top 20 significant downregulated and upregulated genes are highlighted as purple. FDR corrected p-values threshold is 0.01 ($-Log_{10} = 2$). Red: Upregulated, Green: Downregulated, Black: Not significant or low than 2-fold. (**C**) Dysregulated pathways of differentially expressed genes for LUSC low-risk group. (**D**) Dysregulated pathways of differentially expressed genes for LUSC high-risk group.

LUSC signature genes have 10 common genes (EDN1, JHY, PLAAT1, HKDC1, ITIH3, KBTBD11, RGMA, ZNF703, S100A5, LPAL2) with DEGs of both risk groups. Three of the signature genes, ADAMTS17, IGSF1, and LINC01426, are upregulated in the low-risk group; others, NOS1 and SRP14-AS1 are downregulated while Y_RNA is upregulated in the high-risk group.

Risk groups of LUSC have common dysregulated pathways (Figure 5C,D), which are highly related to cancer, such as Cell cycle, DNA replication, Base excision repair, p53 signaling pathway which are upregulated at both risk groups (Figure S19), on the other hand, they also share ECM–receptor interaction, Cell adhesion molecules, Focal adhesion pathways with immune system-related pathways such as Chemokine signaling pathway, Complement and coagulation cascades, Cytokine–cytokine receptor interaction, which are downregulated at both risk groups (Figure S19). However, the high-risk group has more upregulated metabolic pathways such as Central carbon metabolism in cancer, Protein digestion and absorption, Alanine, aspartate and glutamate metabolism, Arginine and proline metabolism, Cysteine and methionine metabolism, Glutathione metabolism, Ribosome biogenesis in eukaryotes; and downregulated immune-related pathways such as JAK-STAT signaling pathway, TNF signaling pathway, Primary immunodeficiency, T cell receptor signaling pathway distinctly from low-risk group (Figure S19). LUSC low-risk group has downregulated PI3K-Akt signaling pathway, Phenylalanine metabolism, Tyrosine metabolism, Phospholipase D signaling pathway, Proteoglycans in cancer and Tight junction pathways with upregulated Hippo signaling pathway and Small-cell lung cancer distinctly from high-risk group (Figure S19).

Active subnetworks of differentially expressed genes in tumor samples of the LUSC risk groups has 357 genes for the low-risk group while 350 genes for high-risk group including 245 common genes (Figure S17). Active pathways of the LUSC risk groups, are highly related to cancer pathways such as PI3K-Akt signaling pathway, Ras signaling pathway, Small-cell lung cancer, Proteoglycans in cancer and Rap1 signaling pathway (Figure 6A,B). LUSC risk groups have mostly similar cancer-related active pathways, however only low-risk group has Nucleotide excision repair, Adherens junction and Alpha-Linolenic acid metabolism pathways, while high-risk group has cancer and metabolism-related pathways such as Basal cell carcinoma, Prolactin signaling pathway, Apoptosis, Mitophagy, Choline metabolism in cancer, Insulin signaling pathway, Carbohydrate digestion and absorption, Central carbon metabolism in cancer with immune system-related Measles and Influenza A pathways.

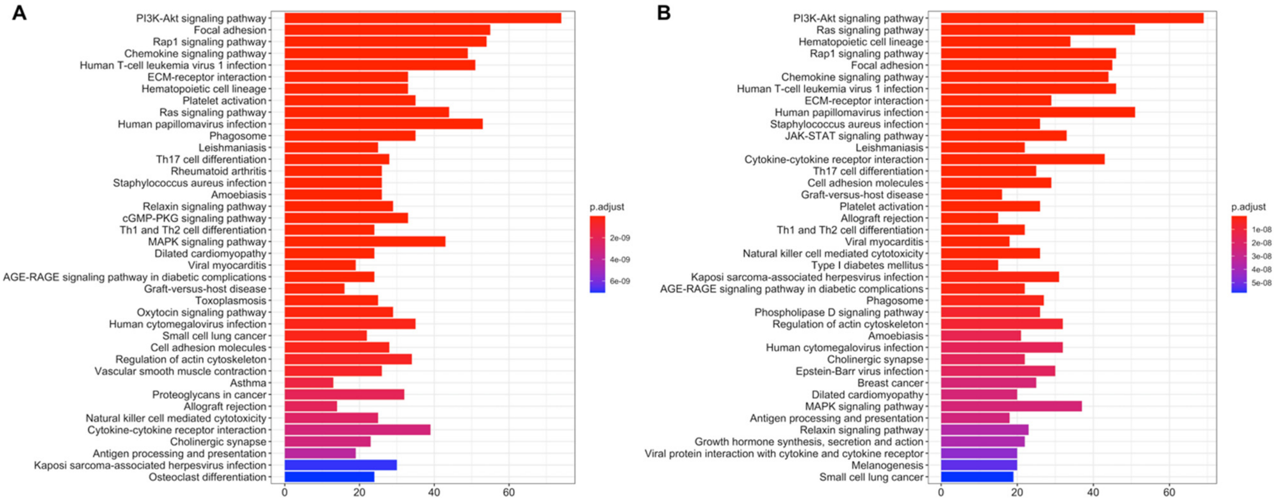


Figure 6. Pathway enrichment of differentially expressed genes at active subnetworks of the LUSC risk groups. Active subnetworks were determined by using differential expression analysis results and pathway enrichment analysis was performed for the genes at subnetworks. (**A**) Active pathways of differentially expressed genes for LUSC low-risk group. (**B**) Active pathways of differentially expressed genes for LUSC high-risk group.

3.3. Copy Number Variations Analysis

The significant aberrant genomic regions in tumor samples of patients were determined and then gene enrichment from genomic regions which have differential copy number was performed. Pathway enrichment analysis of genes which have CNVs was performed and plotted. LUAD low- and high-risk groups have different CNV profiles as seen at CNV plots showing amplified or deleted genomic regions on chromosomes. Chromosomes 1, 6, 7, 10, 13, 16, 17, 28 and 20 have different significant aberrant genomic regions ($q < 0.01$) between risk groups (Figure 7A,B). The highest frequencies of the amplified genes are 45%, 49% and the deleted genes are 31%, 45% in the low- and high-risk groups, respectively. The top 10 the highest frequently amplified or deleted genes in tumor samples of risk groups are different and patients in the same group may have different aberration patterns (Figure 7C,D). The numbers of the deleted genes and the amplified genes are 10,144 and 10,412, respectively, in tumor samples of the LUAD low-risk group. LUAD high-risk group has 5379 deleted and 8442 amplified genes in tumor samples. Risk groups have 4921 deleted and 6559 amplified genes in common (Figure S22).

Pathways of CNV genes are different between LUAD risk groups; mostly immune system pathways such as Allograft rejection, Graft-versus-host disease, Antigen processing and presentation, Complement and coagulation cascades, Inflammatory bowel disease and Viral carcinogenesis pathways have amplified CNVs in the low-risk group (Figure S20) while Herpes simplex virus 1, Cytosolic DNA sensing pathway, Natural killer cell mediated cytotoxicity and Nod-like receptor signaling pathways have deleted CNVs (Figure S20) in the high-risk group (Figure 7). Complement and coagulation cascades pathway has amplified genes in both risk groups while Natural killer cell mediated cytotoxicity and Nod-like receptor signaling pathways have deleted genes in both risk groups (Figure S20). The low-risk group patients have immune system pathways with amplified genes whereas high-risk group patients have immune system pathways with deleted genes. On the other hand, high-risk group has amplified genes in metabolic pathways such as Gastric acid secretion and Insulin secretion (Figure S20).

LUSC risk groups have different significant aberrant genomic regions obviously on chromosomes 5, 6, 8 and X (Figure 8A,B). The highest frequencies of amplified genes are 84%, 77% and of the deleted genes are 55%, 51% in the low- and high-risk groups, respectively. LUSC risk groups have higher frequency of amplified genes than deleted genes. Risk groups have common genes from top 25 the highest frequently amplified genes such as SOX2, GHSR, TNFSF10 and miRNAs, miR-7977 and miR-569, with variable frequencies. Risk groups have also common deleted genes such as CDK inhibitors, CDKN2A and CDKN2B, and miR-1284 (Figure 8C,D). LUSC low-risk group has 10,720 deleted and 10,264 amplified genes while LUSC high-risk group has 9477 deleted and 10,250 amplified genes in tumor samples. Risk groups have 7820 deleted and 8659 amplified genes in common (Figure S22).

Pathways of CNV genes highly overlap between LUSC risk groups and they share cancer-related pathways such as PI3K-Akt signaling pathway, JAK-STAT signaling pathway, Ras signaling pathway, Gastric cancer (Figure 8E,F). However, some pathways differ between risk groups, low-risk group has CNVs at mTOR signaling pathway, VEGF signaling pathways and Central carbon metabolism in cancer, while high-risk group has CNVs at Chemical carcinogenesis, Drug metabolism—cytochrome P450, Carbohydrate digestion and absorption pathways (Figure 8E,F). Steroid hormone biosynthesis and Bile secretion pathways have multiple amplified genes while NOD-like receptor signaling pathway has deleted genes, in both risk groups. Only low-risk group has multiple amplified genes at Growth hormone synthesis, secretion and action, and Complement and coagulation cascades pathways. Only high-risk group has amplified genes at Chemical carcinogenesis and Drug metabolism pathways while has deleted genes at Cytokine-cytokine receptor interaction and Fatty acid biosynthesis pathways (Figure S21).

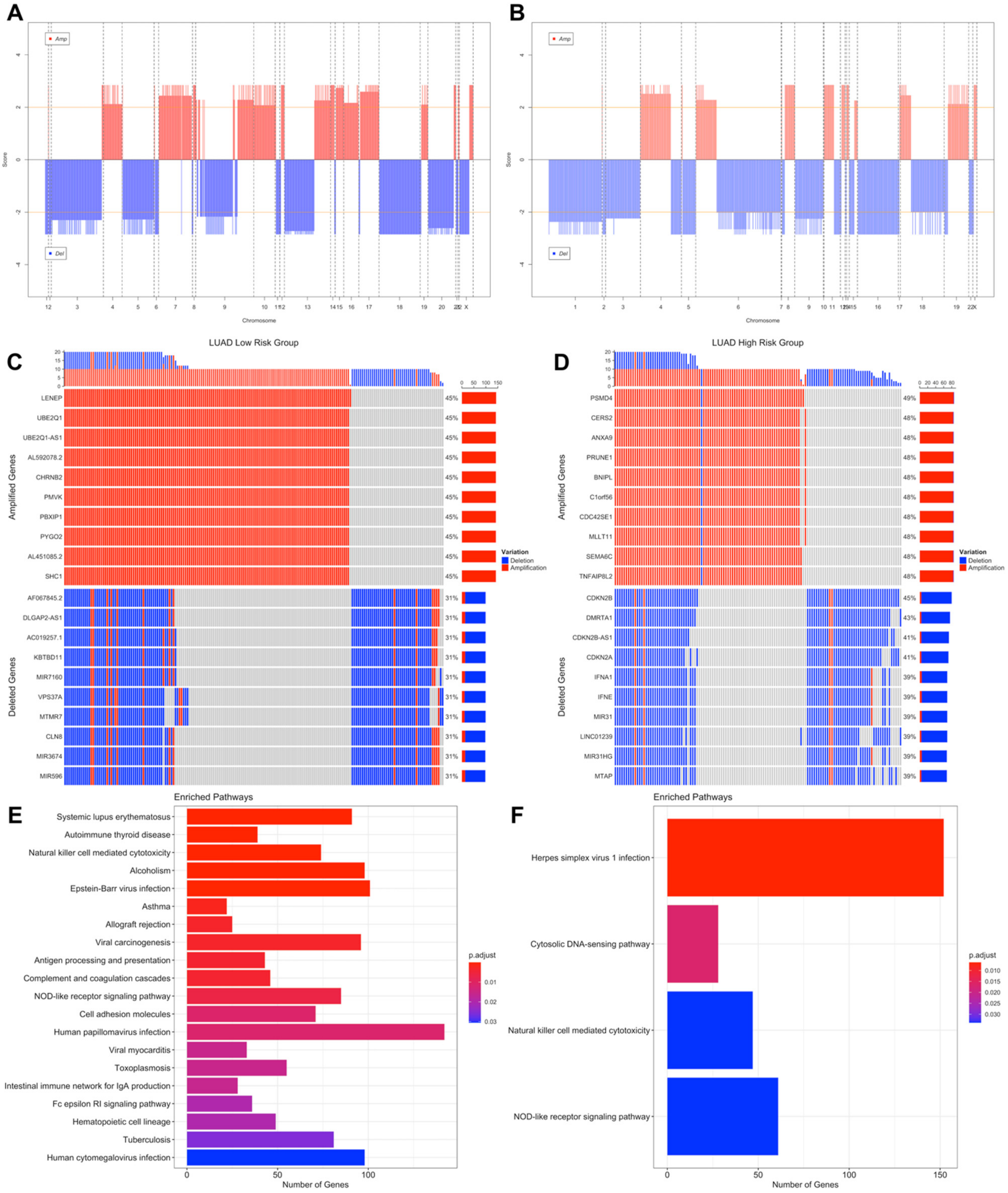


Figure 7. Significant Copy Number Variations (CNVs) of the LUAD risk groups. (**A**) CNV plot at genome scale showing amplified or deleted genomic regions on chromosomes of the LUAD low-risk group. Score: -Log$_{10}$(q value), Horizontal orange line: 0.01 q value threshold. (**B**) CNV plot of the LUAD high-risk group. (**C**) OncoPrint plot showing 25 the highest frequently amplified and deleted genes of the LUAD low-risk group. (**D**) OncoPrint plot showing 25 the highest frequently amplified and deleted genes of the LUAD high-risk group. (**E**) Pathways of CNV genes of the LUAD low-risk group. (**F**) Pathways of CNV genes of the LUAD high-risk group.

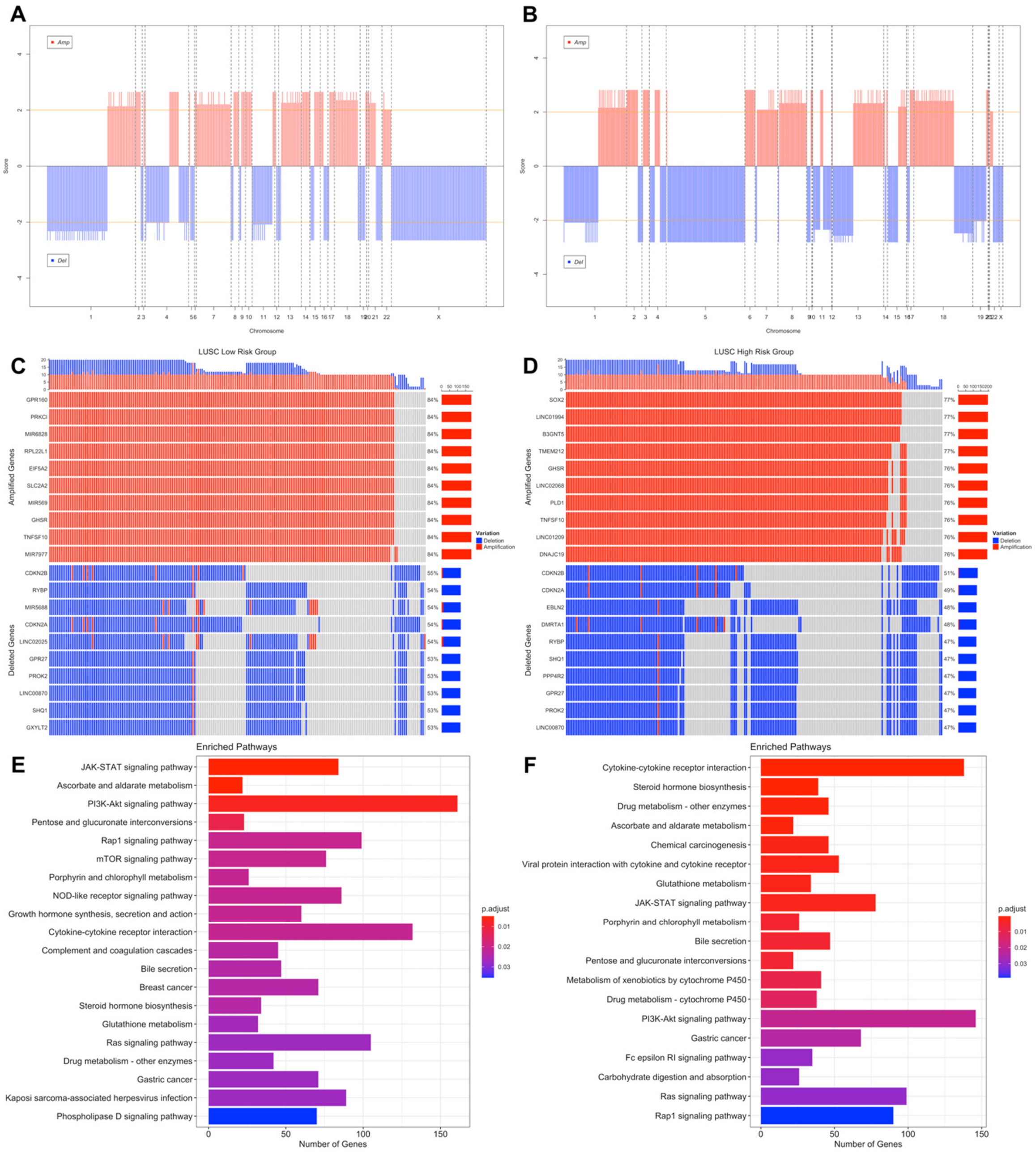


Figure 8. Significant Copy Number Variations (CNVs) of the LUSC risk groups. (**A**) CNV plot at genome-scale showing amplified or deleted genomic regions on chromosomes of the LUSC low-risk group. (**B**) CNV plot of the LUSC high-risk group. (**C**) OncoPrint plot showing 25 the highest frequently amplified and deleted genes of the LUSC low-risk group. (**D**) OncoPrint plot showing 25 the highest frequently amplified and deleted genes of the LUSC high-risk group. (**E**) Pathways of CNV genes of the LUSC low-risk group. (**F**) Pathways of CNV genes of the LUSC high-risk group.

3.4. *Simple Nucleotide Variations Analysis*

Significantly ($q < 0.05$) mutated genes classified as oncogene (OG) or tumor suppressor gene (TSG) based on TSG/OG scores of the genes and the Cancer Gene Census, were identified for LUAD and LUSC risk groups. COSMIC database was used as a reference mutation database for this analysis and Cancer Gene Census data.

LUAD low-risk group has 15,376 mutated genes, while LUAD low-risk group has 12,815 mutated genes, 11,516 genes of which are common between LUAD risk groups (Figure S27). LUAD patients have a wide range of mutation numbers changing from

1518/1158 to 10s with median 167 and 172.5 for low- and high-risk groups, respectively. Missense mutation is the highest frequent mutation type, and C > A and C > T substitutions are the most frequent ones for both risk groups. LUAD risk groups have a similar set of mutated genes with varying frequencies. TP53 is the highest frequently mutated gene with 45% and 53% for low- and high-risk groups, and the following ones are MUC16 (39%, 40%) and CSMD3 (38%, 35%) for both groups (Figure S23). SomInaClust analysis was performed to determine driver genes, and 39 genes and 19 genes are strong candidate driver genes for the low-risk group and high-risk group, respectively (Tables S5 and S6). Interestingly, LUAD risk groups share 18 of these driver genes (Figure S27). SomInaClust calculates TSG and OG scores based on background mutation rate and hot spots, then classifies the genes based on TSG/OG scores and cancer gene census data (Figure S25). The driver genes determined in LUAD low-risk group are KRAS, TP53, EGFR, BRAF, STK11, MGA, NF1, RB1, PIK3CA, ATM, RBM10, SETD2, ARID1A, CTNNB1, CMTR2, SF3B1, CSMD3, ATF7IP, KEAP1, HMCN1, EPHA5, ARID2, TTK, SMAD4, KDM5C, SMARCA4, APC, NFE2L2, RIT1, DDX10, LTN1, CDH10, SPTA1, LRP1B, COL11A1, MAP3K12, USH2A, AKAP6 and RASA1. The driver genes determined in LUAD high-risk group are KRAS, TP53, STK11, EGFR, BRAF, RBM10, PIK3CA, SETD2, ARID2, NF1, RB1, MGA, KEAP1, CSMD3, SMARCA4, CTNNB1, KDM5C, IDH1 and ATM (Figure S25; Tables S5 and S6). TP53 and CSMD3 genes are the most frequently mutated genes with 47%, 56% and 41%, 37% frequencies, respectively for low- and high-risk groups (Figure 9A,B). More than half of the genes are mutated in less than 12% of patients. For common genes, LUAD high-risk group has mostly higher frequencies. TP53 has differential mutation types, while KRAS has mostly missense mutations. CSMD3 has more multi-hits (multiple mutations in one patient) in the low-risk group than the high-risk group. EGFR has in frame deletions in both risk groups and other common genes have similar mutation type pattern between risk groups (Figure 9A,B). Pathways of driver mutated genes are highly lung cancer-related pathways such as Non-small-cell lung cancer, EGFR tyrosine kinase inhibitor resistance, Platinum drug resistance, MAPK signaling, mTOR signaling, Ras signaling pathway, PI3K-Akt signaling (Figure 9C,D) and other immunologic and metabolic pathways such as Signaling pathways regulating pluripotency of stem cells, FoxO signaling pathway, Rap1 signaling pathway, Central carbon metabolism in cancer, Proteoglycans in cancer, Human T-cell leukemia virus 1 infection, PD-L1 expression and PD-1 checkpoint pathway in cancer and Natural killer cell mediated cytotoxicity pathways, for both risk groups. Many common pathways are enriched because these mutated driver genes play role in many crucial important pathways. However, Wnt signaling pathway and Hippo signaling pathways are mutated only in the low-risk group, while Gap junction, GnRH signaling pathway, C-type lectin receptor signaling pathway, T cell receptor signaling pathway, HIF-1 signaling pathway, Growth hormone synthesis, secretion and action and AMPK signaling pathways are mutated only in the high-risk group (Figure 9C,D).

LUSC low-risk group has 14,038 mutated genes, while LUSC low-risk group has 14,616 mutated genes, and 11,947 genes are common (Figure S27). LUSC patients have a range of mutation numbers from 2300/1488 to 10s with median 201 for low- and high-risk groups, respectively. Missense mutation is the highest frequent mutation type, and C > A and C > T substitutions are the most frequent ones for both risk groups. LUSC risk groups have overlapping list of mutated genes with varying frequencies. TP53 is the highest frequently mutated gene with 80% and 78% for low- and high-risk groups, and the following ones are CSMD3 (42%, 42%) and MUC16 (39%, 40%) for both groups (Figure S24). As candidate driver genes, 30 genes and 19 genes were identified for the low-risk group and the high-risk group, respectively (Tables S7 and S8). LUSC risk groups share 14 of these driver genes (Figure S27). The driver genes determined in LUSC low-risk group are TP53, KMT2D, NFE2L2, PIK3CA, CDKN2A, PTEN, RB1, FAT1, ARID1A, NF1, RASA1, CUL3, KDM6A, NRAS, KRT5, ZNF750, EP300, FGFR3, TAOK1, CSMD3, NSD1, HRAS, SI, PDS5B, KRAS, KEAP1, API5, HNRNPUL1, SLC16A1, FBXW7. The driver genes determined in LUSC high-risk group are TP53, NFE2L2, PIK3CA, KMT2D, FAT1, CDKN2A, RB1, PTEN, NOTCH1, ARID1A, RASA1, NF1, KMT2C, BRAF, PIK3R1, CSMD3, STK11, HRAS, KEAP1 (Figure S26; Tables S7 and S8). TP53 (83%, 82%), CSMD3 (44%, 44%) and KMT2D (25%, 23%) are most frequent mutated genes for low- and high-risk groups (Figure 10A,B). For common genes, risk groups have similar frequencies. TP53 and KMT2D genes have differential mutation types, while CSMD3 has mostly missense and multi-hit mutations. CDKN2A has mostly truncating mutations in both risk groups and other common genes have similar mutation type pattern between risk groups (Figure 10A,B). Pathways of driver mutated genes are highly lung cancer-related pathways such as Non-small-cell lung cancer, EGFR tyrosine kinase inhibitor resistance, Platinum drug resistance, MAPK signaling and Ras signaling (Figure 10C,D) and other immunologic and metabolic pathways such as FoxO signaling pathway, Central carbon metabolism in cancer, Proteoglycans in cancer, Hepatitis B, Hepatitis C, PD-L1 expression and PD-1 checkpoint pathway in cancer for both risk groups. Many common pathways are enriched because these mutated driver genes play role in many crucial important pathways. However, Gap junction and Ubiquitin mediated proteolysis pathways are mutated only in the low-risk group, while HIF-1 signaling and TNF signaling pathways are mutated only in the high-risk group (Figure 10C,D).

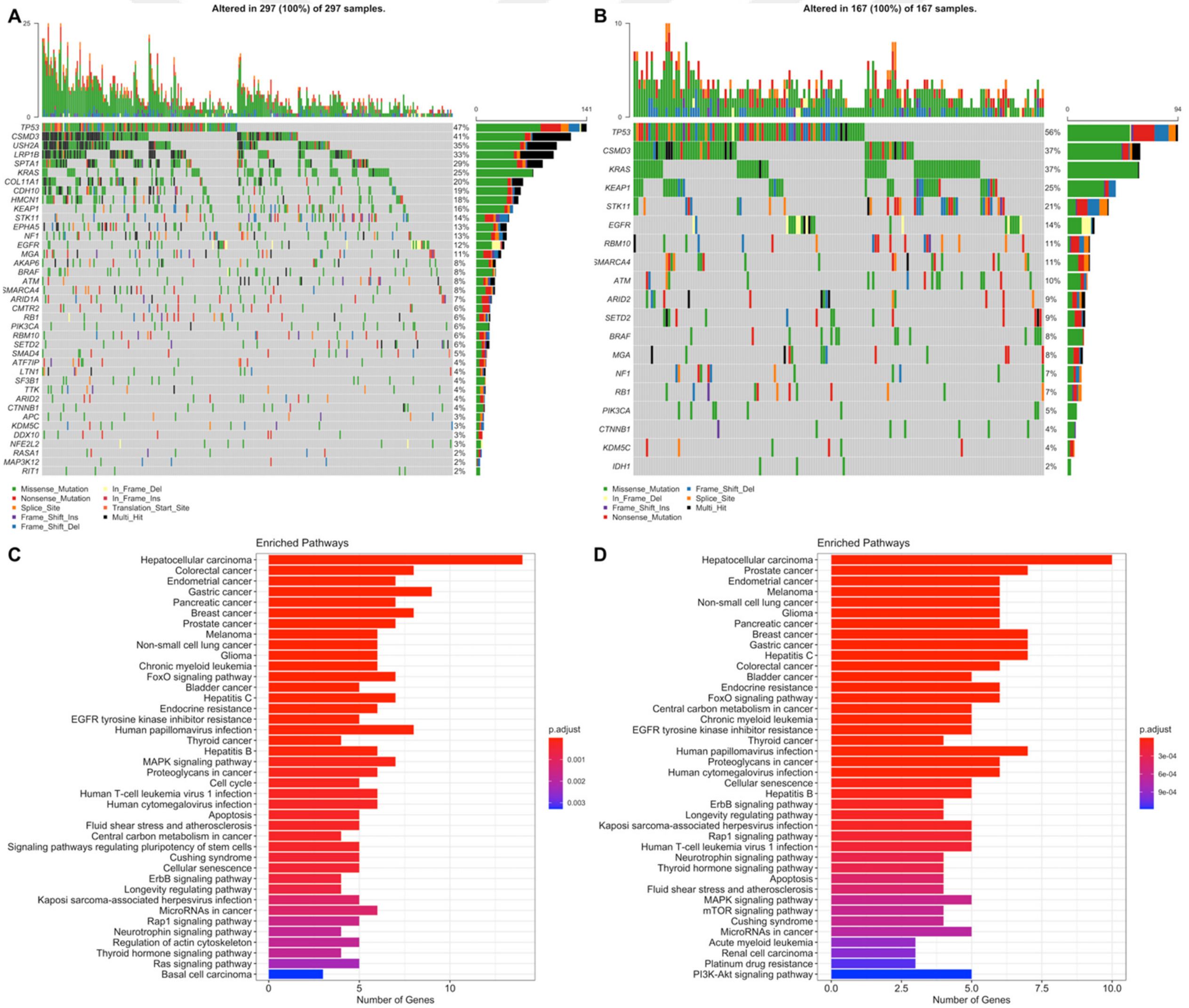


Figure 9. Oncoplot of potential driver genes containing significant SNVs of the LUAD risk groups. (**A**) Oncoplot showing significant SNV genes in tumor samples of the LUAD low-risk group patients. (**B**) Oncoplot showing significant SNV genes in tumor samples of the LUAD high-risk group patients. (**C**) Pathway enrichment of the significant SNV genes of the LUAD low-risk group. (**D**) Pathway enrichment of the significant SNV genes of the LUAD high-risk group.

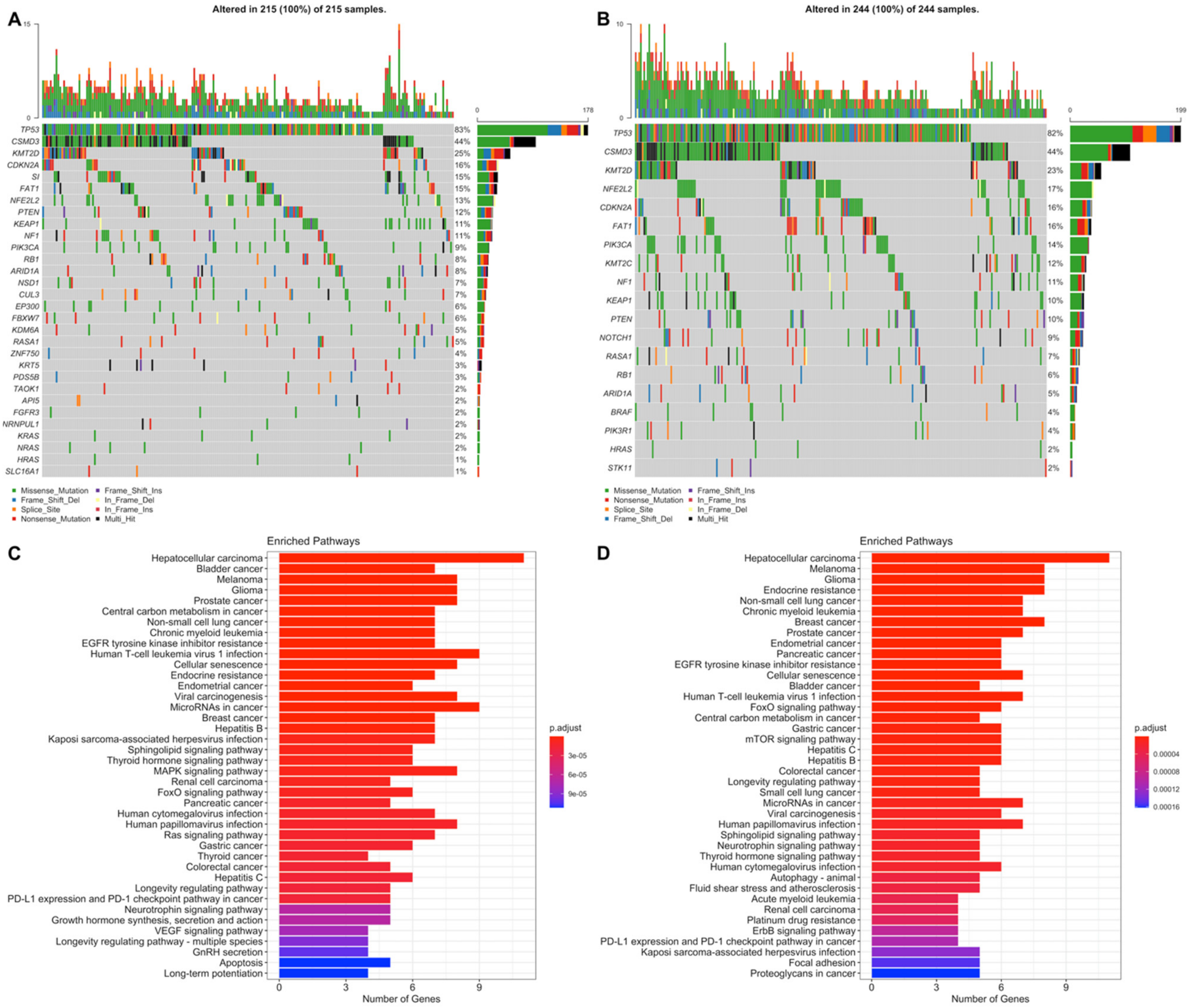


Figure 10. Oncoplot of potential driver genes containing significant SNVs of the LUSC risk groups. (**A**) Oncoplot showing significant SNV genes in tumor samples of the LUSC low-risk group patients. (**B**) Oncoplot showing significant SNV genes in tumor samples of the LUSC high-risk group patients. (**C**) Pathway enrichment of the significant SNV genes of the LUSC low-risk group. (**D**) Pathway enrichment of the significant SNV genes of the LUSC high-risk group.

When venn diagram is drawn by using all driver genes, all cancer and risk groups have TP53, CSMD3, KEAP1, NF1, RB1 and PIK3CA mutations. KRAS, STK11, BRAF, ARID1A, NFE2L2 and RASA1 genes are shared by 3 different groups. LUAD high-risk group has only IDH1 oncogene as different from LUAD low-risk group while LUSC high-risk group has KMT2C, NOTCH1 and PIK3R1 tumor suppressor genes as different from LUSC low-risk group. EGFR, MGA and SMARCA4 are not driver genes in LUSC while CDKN2A, PTEN, HRAS and FAT1 are not driver genes in LUAD groups (Figure 11).

Significant SNVs and CNVs on driver genes are co-displayed as OncoPrint. Although there exist some genes with both SNVs and significant CNVs while others have only SNVs. Moreover, some patients have only SNVs or only CNVs or both for a particular driver gene.

TP53, STK11, KEAP1, SMARCA4 and MGA genes have deletions while CSMD3 and PIK3CA genes have amplification beside SNVs in both LUAD risk group. KRAS and EGFR genes have amplification in the high-risk group; however, they do not have significant CNVs in the low-risk group. Oncogenes tend to have amplifications while tumor suppressor genes tend to have deletions in both risk groups with exceptions (CSMD3, CDH10, HMCN1, AKAP6 and CTNNB1) (Figure 12).

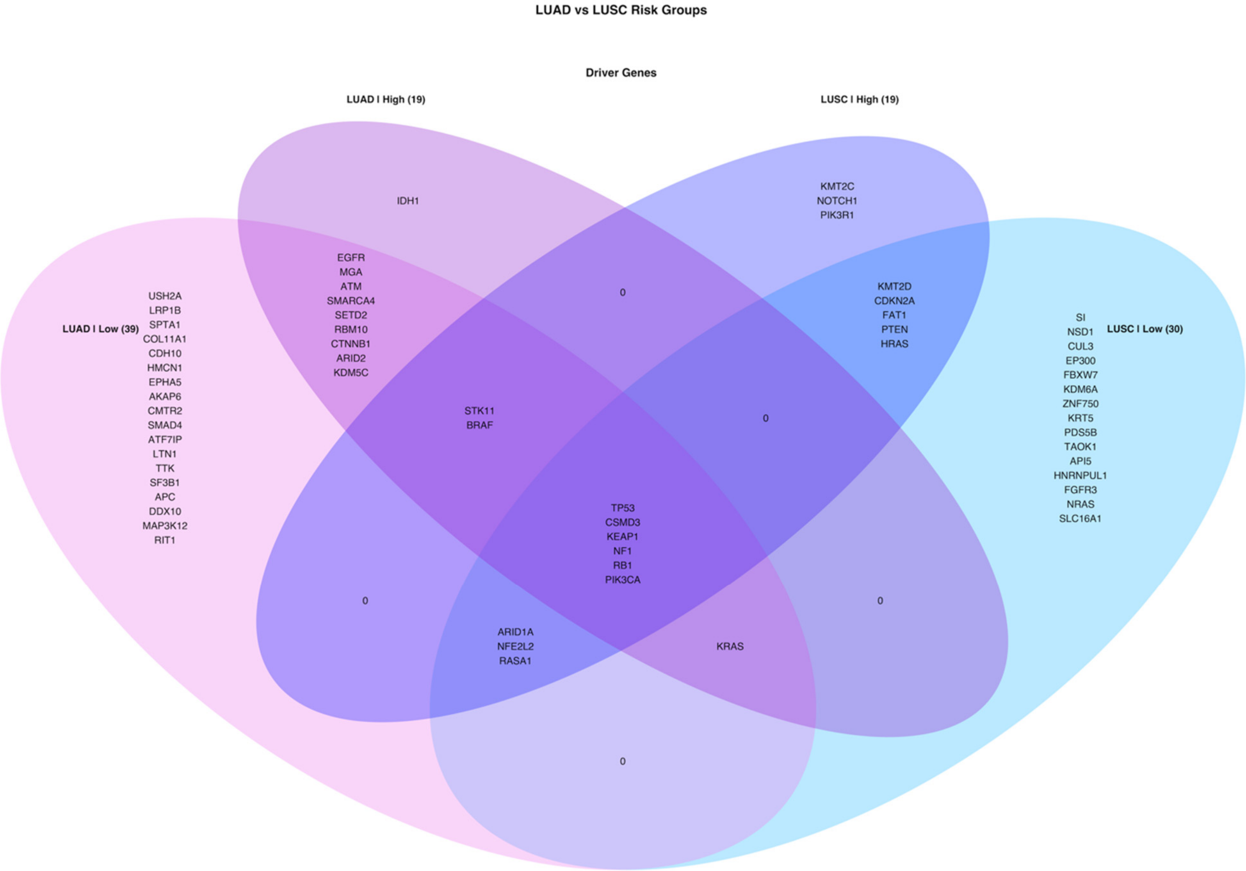


Figure 11. Venn diagram of driver genes containing Simple Nucleotide Variation (SNV) in tumor samples of LUAD and LUSC risk groups.

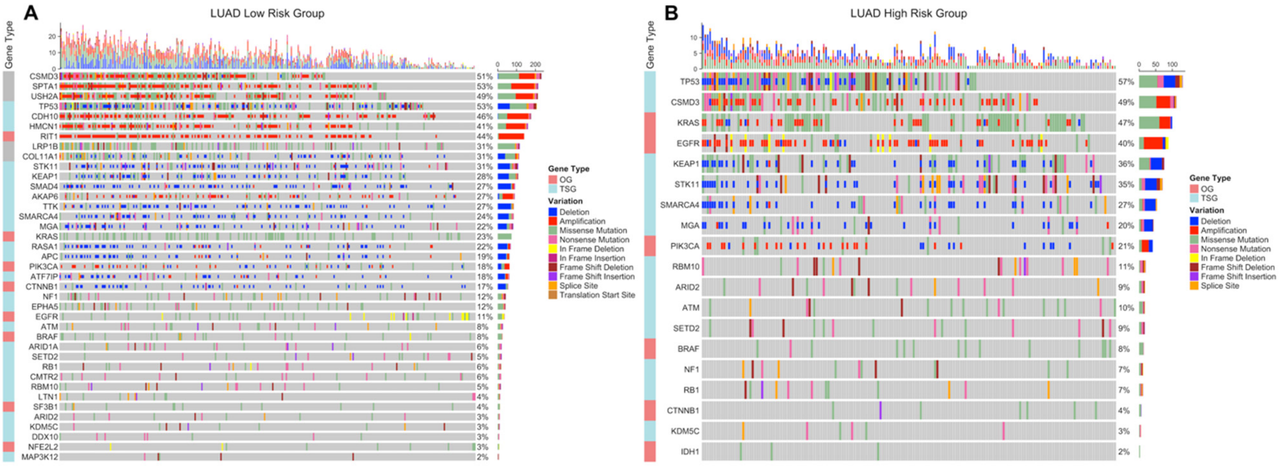


Figure 12. OncoPrint of the driver genes containing significant SNVs and CNVs in LUAD risk groups. Significant SNVs and CNVs are plotted together on potential driver genes in tumor samples of the LUAD risk groups. (**A**) OncoPrint of the driver genes in LUAD low-risk group. (**B**) OncoPrint of the driver genes in LUAD high-risk group.

OncoPrints in Figure 13 show that TP53, CDKN2A, FAT1, RASA1, ARID1A and HRAS genes have deletions while only PIK3CA gene has amplification beside SNVs in both LUSC risk groups. PIK3R1, KEAP1 and STK11 genes have deletions only in the high-risk group while SI, CSMD3, ZNF750, KRAS genes have amplification and NSD1, FGFR3, PTEN, SLC16A1, NRAS and CUL3 have deletion only in the low-risk group. Oncogenes tend to have amplifications while tumor suppressor genes tend to have deletions in both risk groups with exceptions (CSMD3, FGFR3, ZNF750, NRAS, HRAS, KEAP1) (Figure 13).

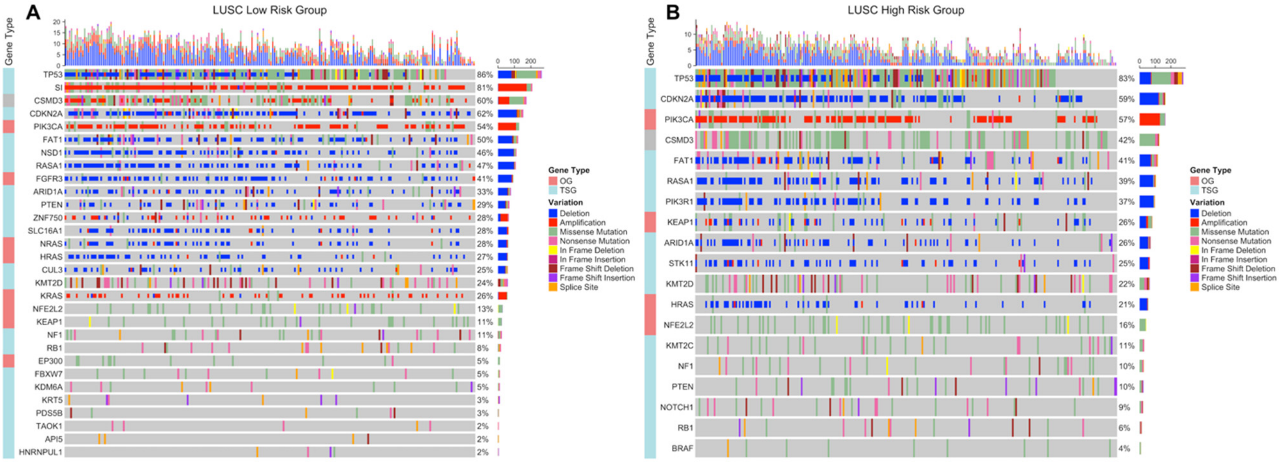


Figure 13. OncoPrint of the driver genes containing significant SNVs and CNVs in LUSC risk groups. Significant SNVs and CNVs are plotted together on potential driver genes in tumor samples of the LUSC risk groups. (**A**) OncoPrint of the driver genes in LUSC low-risk group. (**B**) OncoPrint of the driver genes in LUSC high-risk group.

Circos plots showing all non-synonymous SNVs in original data of risk groups and significant CNVs at genomic scale on chromosomes were drawn to show the genomic alterations between risk groups of LUAD and LUSC.

LUAD low-risk group has more genome-wide CNVs and SNVs than the high-risk group. The low-risk group has more genomics regions containing missense, nonsense and frame-shift insertions/deletions mutations. Moreover, low-risk group has extra deletions on chromosomes 1, 3, 5, 6, 12, 15 and X with extra amplifications on chromosomes 6, 10, 14, and 20. The high-risk group has extra amplifications on chromosomes 7, 11, 12, and 17. The CNVs of high-risk group are localized mostly on 1, 3, 5, 6, 7, 8 and 17 whereas low-risk group has CNVs on more chromosomes (Figure 14).

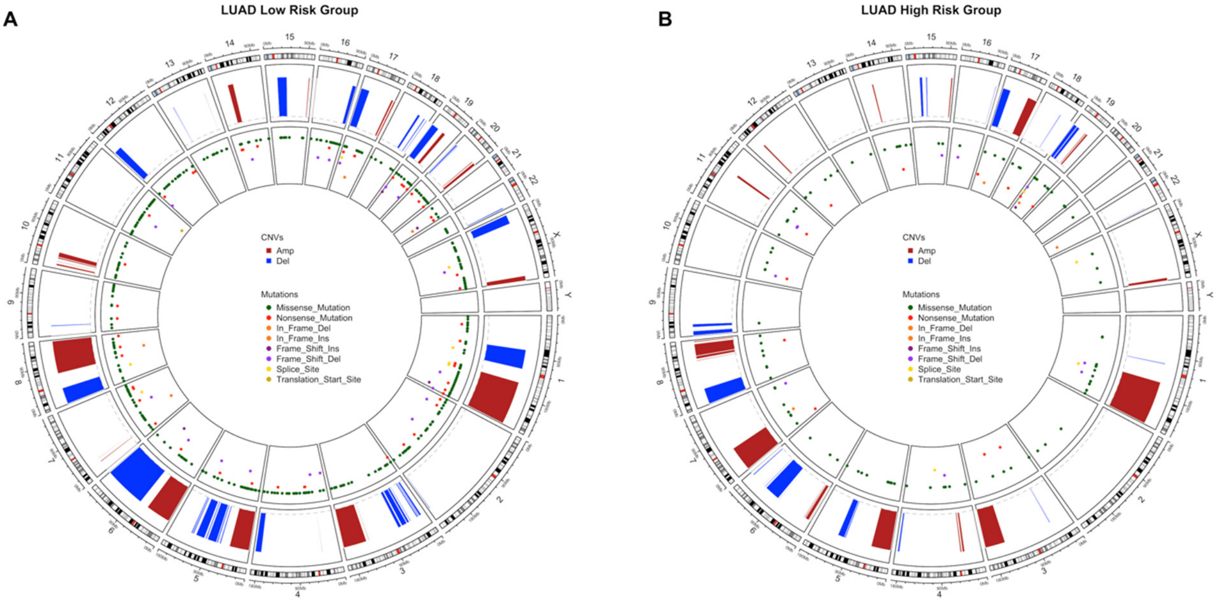


Figure 14. Circos plot of chromosome regions containing all SNVs and CNVs in LUAD risk groups. Significant CNVs ($q < 0.01$) and all SNVs in original data are plotted together on chromosome regions in tumor samples of the LUAD risk groups. (**A**) Circos plot of the LUAD low-risk group. (**B**) Circos plot of the LUAD high-risk group.

LUSC high-risk group has more genomic regions containing missense and nonsense mutations than the low-risk group. However, they have similar amount of CNVs although with different localizations. The high-risk group has extra amplifications on chromosomes 4, 6 and 11; has extra deletions on chromosomes 15, 19 and X. The low-risk group has only extra deletions on chromosomes 1, 5, 6, 11 and 16 (Figure 15).

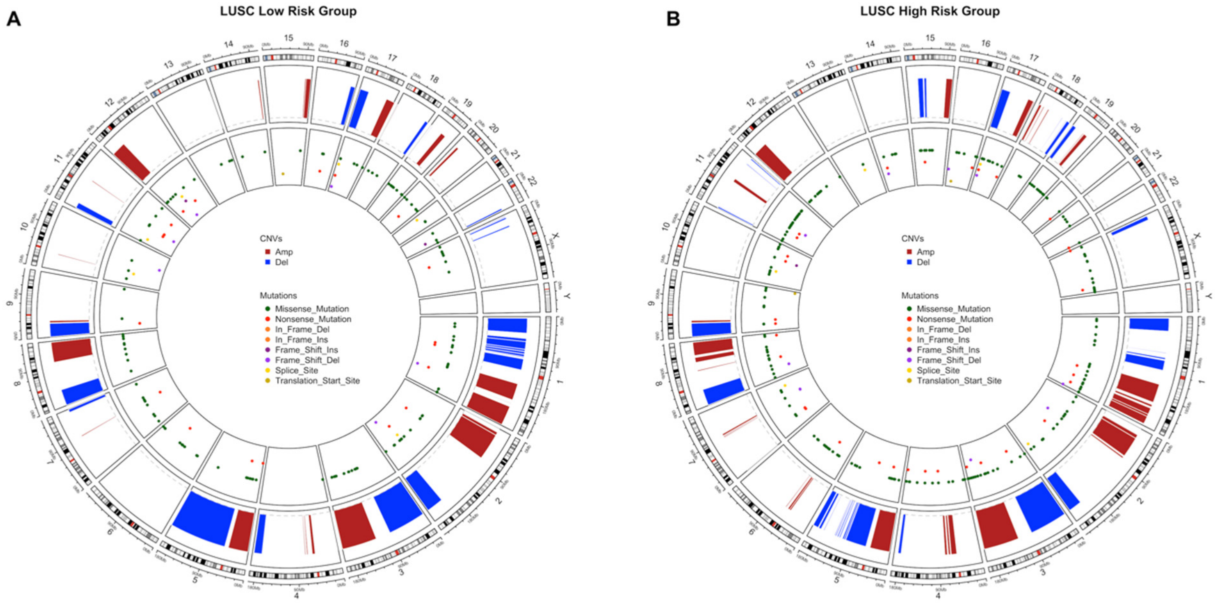


Figure 15. Circos plot of chromosome regions containing all SNVs and CNVs in LUSC risk groups. Significant CNVs ($q < 0.01$) and all SNVs in original data are plotted together on chromosome regions in tumor samples of the LUSC risk groups. (**A**) Circos plot of the LUSC low-risk group. (**B**) Circos plot of the LUSC high-risk group.

4. Discussion

In order to profile the genetic differences between risk groups of LUAD and LUSC, gene expression signatures were generated and the patients were clustered into low- and high-risk groups and then significant DEGs, DEGs at active subnetworks, CNVs and SNVs were identified in each risk group. The biological alterations for these data types were compared between risk groups and between lung cancer subtypes.

Expression signature for LUAD consists of 35 gene which 27 of are protein-coding genes while two are long intergenic non-protein coding RNA, one is antisense RNA, three are pseudogenes and two are novel transcripts. Many of the coding genes are lung cancer or other cancer types related such as ADAMTS15 [44], ASB2 [45] and EPHX1 [46] with potential tumor suppressor roles; ANGPTL4 [47], ASCL2 [48], CCL20 [49], DKK1 [50], GRIK2 [51], LDHA [52], RGS20 [53], RHOQ [54], TLE1 [55] and WBP2 [56] with potential oncogenic roles; and CD200 [57], CD200R1 [57], CCDC181 [58], GNPNAT1 [59], IRX2 [60], LDLRAD3 [61], STAP1 [62], LINC00578 [63] with prognostic potential. Moreover, MS4A1 is dysregulated in asbestos-related lung squamous carcinoma [64], RAB9B is a target of miR-15/16 which are highly related to lung cancer [65], LINC00539 is related to tumor immune response [66] while long non-coding RNA, OGFRP1, regulates non-small-cell lung cancer progression [67]. The remaining signature genes, CPXM2, ENPP5, SAMD13, SLC52A1, ZNF682, ZNF835, ZNF571-AS1 and U91328.1, have not been related to carcinoma, yet. However, they showed highly prognostic power through risk score to distinguish low- and high-risk of overall survival in LUAD.

LUSC gene expression signature including 33 genes of which ALDH7A1 [68], ALK [69], EDN1 [70], FABP6 [71], HKDC1 [72], IGSF1 [73], KBTBD11 [74], NOS1 [75], SLC9A9 [76], STK24 [77], UBB [78], ZNF703 [79] have been shown with oncogenic relations while RGMA [80] is candidate tumor suppressors. ITIH3 [81] and S100A5 [82] has been related to prognostic biomarker potentials. Other cancer-related genes are ADAMTS17 [83], LINC01748 [84], LPAL2 [85], SRP14-AS1 [86] and WASH8P [87]. Long intergenic non-protein coding RNA, LINC01426, promotes cancer progression via AZGP1 and predicts poor prognosis in patients with LUAD [88]. COL28A1 has prognostic values in glioblastoma [89]. Many of the genes such as JHY, PLAAT1, PNMA8B, RPL37P6, SNX32, UGGT2 and Y_RNA have not been related to any cancer, yet.

Gene expression signatures of LUAD and LUSC share eight pathways which are mostly metabolic pathways. LUAD signature plays role in immune-related pathways as different from those in LUSC. However, pathway enrichment shows us that risk prediction works on metabolic pathways, therefore if we put a name to important mutations as driver mutations, in this case we can say that reprogramming of energy metabolism is the alternative fuel of the cancer [90–92]. The differential expression on them with immune system effect in count can hold the passage of cancer.

High-risk groups of both LUAD and LUSC have more immune pathways including downregulated genes and metabolic pathways including upregulated genes. On the other hand, low-risk groups have both upregulated and downregulated genes on cancer-related pathways. Although LUAD and LUSC seem to have similar characteristics of risk groups, close signature gene pathways and similar differential expression pathways sharing 2106 DEGs in total, they are displayed separately in PCA, especially at analysis of test groups.

At CNV level both risk groups and cancer subtypes have huge number of genes with amplifications or deletions which can cause genomic instability and uncontrolled regulation. Both LUAD and LUSC risk groups have important gene alterations such as CDKN2A and CDKN2B deletions which are associated with NSCLC [93] and promotes KRAS and EGFR mutant tumorigenesis [94,95] while SOX2 oncogene amplification in LUSC which is a common event in squamous cell carcinomas [96,97] and amplification of PSMD4 in LUAD, with oncogenic roles in breast, hepatocellular, colorectal and prostate cancer cells [98–101]. CNVs also play role in metabolic and immune-related pathways which can differ between risk groups and cancer subtypes. If we look from a higher perspective, the LUAD low-risk group has much more CNVs and SNVs on its genome than the high-risk group. On the other hand, the LUSC high-risk group has more SNVs than the low-risk group while CNVs do not vary too much.

SNV analysis gives similar results with literature for example EGFR and KRAS mutations are mutually exclusive in LUAD samples that is confirmed again [9]. Additionally, EGFR [102], MGA [103], SMARCA4 [104], ATM [105], RBM10 [106] and KDM5C [107] which are lung cancer related genes are mutated only in LUAD but not in LUSC. On the other hand, CDKN2A [108], PTEN [109] and HRAS [110] genes are mutated only in LUSC. In general, low-risk groups have more mutated genes for both LUAD and LUSC samples. When SNV and CNV genes are plotted together, it can be seen that LUAD high-risk group has obvious oncogene amplifications and tumor suppressor deletions, while LUAD low-risk group has both tumor suppressor deletions and tumor suppressor amplifications with a few oncogene amplifications. This SNV and copy number differential pattern can cause differential gene expression profiles and characteristics of tumor. LUSC patients have mostly deletions on driver genes with only PIK3CA [111] and KRAS [111] oncogene amplifications. Both LUSC risk groups have obvious TP53 [111] and CDKN2A tumor suppressor gene deletions, but amplification of CSMD3, which has differential roles in lung cancer [112,113], does not occur in LUSC high-risk group. Again, only these driver genes which have differential alterations and frequencies can create the risk difference based on gene expression levels.

5. Conclusions

This study has been performed to profile the genomic and transcriptomic differences not only between LUAD and LUSC but also between risk groups to understand the driving differences between them. Treatment options can vary between cancer subtypes and risk groups because of differential targetable mutation patterns. Nowadays, many groups and government institutions are working on the integration of the drug bioactivity and molecular data to investigate more effective molecularly targeting therapeutics for individual patients for the personalized therapy.

Supplementary Materials: The supplementary data are available online at https://www.mdpi.com/2075-4426/11/2/154/s1; Figure S1: Flowchart of method and used R packages in this study. The other R packages not written in this flowchart can be found at Materials and Method part of the article; Figure S2: Gene expression signature and risk clustering of LUAD training dataset; Figure S3: Survival analysis of risk groups clustered by using signature gene expression at different tumor stages in LUAD training dataset; Figure S4: Mosaic plots showing association analysis of categorical variables for LUAD training dataset. Pearson residuals show the positive (blue) or negative (red) association between levels of categories; Figure S5: Multivariate Cox Regression results of clinical variables and risk score in LUAD training dataset. Only risk score has significant result when all clinical variables are included into multivariate analysis; Figure S6: Multivariate Cox Regression results of selected clinical variables (which have significant results in univariate Cox analysis) and risk score in LUAD training dataset. Risk score, t, n, m stages and history of prior malignancy have significant effects on survival. When pathologic tumor stage is used instead of t, n, m stages, only risk score and history of prior malignancy show significant effect on survival; Figure S7: Survival analysis of risk groups clustered by using signature gene expression at different tumor stages in LUAD test dataset; Figure S8: Mosaic plots showing association analysis of categorical variables for LUAD test dataset; Figure S9: Multivariate Cox Regression results of selected clinical variables (which have significant results in univariate Cox analysis) and risk score in LUAD test dataset. Risk score and n stages have significant effect on survival. When pathologic tumor stage is used instead of t, n, m stages, only risk score shows significant effect on survival; Figure S10: Gene expression signature and risk clustering of LUSC training dataset; Figure S11: Survival analysis of risk groups clustered by using signature gene expression at different tumor stages in LUSC training dataset; Figure S12: Mosaic plots showing association analysis of categorical variables for LUSC training dataset. Pearson residuals show the positive (blue) or negative (red) association between levels of categories; Figure S13: Multivariate Cox Regression results of selected clinical variables (which have significant results in univariate Cox analysis) and risk score in LUSC training dataset. Risk score, tissue or organ of origin, t and n stages and history of prior malignancy have significant effects on survival. When pathologic tumor stage is used instead of t, n, m stages, tissue or organ of origin, risk score and history of prior malignancy show significant effect on survival; Figure S14: Survival analysis of risk groups clustered by using signature gene expression at different tumor stages in LUSC test dataset; Figure S15: Mosaic plots showing association analysis of categorical variables for LUSC test dataset. Pearson residuals show the positive (blue) or negative (red) association between levels of categories; Figure S16: Multivariate Cox Regression results of selected clinical variables (which have significant results in univariate Cox analysis) and risk score in LUSC test dataset. Only risk score has significant effect on survival either t, n, m stages or pathologic tumor stage is used instead of t, n, m stages; Figure S17: Venn diagram of differentially expressed genes in tumor samples of risk groups for LUAD and LUSC test groups; Figure S18: Pathway enrichment of DEGs of LUAD risk groups; Figure S19: Pathway enrichment of DEGs of LUSC risk groups; Figure S20: Pathway enrichment of CNV genes of LUAD risk groups; Figure S21: Pathway enrichment of CNV genes of LUSC risk groups; Figure S22: Venn diagram of genes which have significant copy number alterations in tumor samples of LUAD and LUSC risk groups; Figure S23: Summary of SNVs in LUAD risk groups; Figure S24: Summary of SNVs in LUSC risk groups; Figure S25: SomInaClust result of potential driver genes containing significant SNVs in LUAD risk groups. SomInaClust calculates oncogene (OG) score and tumor suppressor gene (TSG) score for each significant gene and classifies the gene according to the score threshold (20) and reference database; Figure S26: SomInaClust result of potential driver genes containing significant SNVs in LUSC risk groups. SomInaClust calculates oncogene (OG) score and tumor suppressor gene (TSG) score for each significant gene and classifies the gene according to the score threshold (20) and reference database; Figure S27: Venn diagram of all genes and potential driver genes containing SNVs of LUAD and LUSC risk groups, Table S1: Gene list of expression signature in LUAD. Ensemble Gene IDs were used in signature analysis and then enriched by using BioMart database; Table S2: KEGG pathway enrichment of expression signature gene list in LUAD by using KEGG Mapper tool; Table S3: Gene list of expression signature in LUSC. Ensemble Gene IDs were used in signature analysis and then enriched by using BioMart database; Table S4: KEGG pathway enrichment of expression signature gene list in LUSC by using *clusterProfiler*

R package; Table S5: SomInaClust result of SNV data in tumor samples of LUAD low-risk group; Table S6: SomInaClust result of SNV data in tumor samples of LUAD high-risk group; Table S7: SomInaClust result of SNV data in tumor samples of LUSC low-risk group; Table S8: SomInaClust result of SNV data in tumor samples of LUSC high-risk group.

Author Contributions: Methodology, T.Z.; formal analysis, T.Z.; resources, T.Z., T.Ö.-S.; data curation, T.Z.; writing—original draft preparation, T.Z.; writing—review and editing, T.Ö.-S.; visualization, T.Z.; project administration, T.Ö.-S. All authors have read and agreed to the published version of the manuscript.

Funding: T.Z. and T.Ö.-S. were partially funded by Turkish National Institutes of Health (TÜSEB) grant number 4583.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets supporting the conclusions of this article are publicly available and can be downloaded from TCGA data portal (https://portal.gdc.cancer.gov) or by using *TCGAbiolinks* R package [18]. The R code used in this study is available upon request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. GLOBOCAN 2020: Cancer Today. Available online: https://gco.iarc.fr/today/home (accessed on 29 December 2020).
2. Alexander, M.; Kim, S.Y.; Cheng, H. Update 2020: Management of Non-Small Cell Lung Cancer. *Lung* **2020**, *198*, 897–907. [CrossRef] [PubMed]
3. Chansky, K.; Detterbeck, F.C.; Nicholson, A.G.; Rusch, V.W.; Vallières, E.; Groome, P.; Kennedy, C.; Krasnik, M.; Peake, M.; Shemanski, L.; et al. The IASLC Lung Cancer Staging Project: External Validation of the Revision of the TNM Stage Groupings in the Eighth Edition of the TNM Classification of Lung Cancer. *J. Thorac. Oncol.* **2017**, *12*, 1109–1121. [CrossRef]
4. Camidge, D.R.; Doebele, R.C.; Kerr, K.M. Comparing and contrasting predictive biomarkers for immunotherapy and targeted therapy of NSCLC. *Nat. Rev. Clin. Oncol.* **2019**, *16*, 341–355. [CrossRef] [PubMed]
5. Wang, B.-Y.; Huang, J.-Y.; Chen, H.-C.; Lin, C.-H.; Lin, S.-H.; Hung, W.-H.; Cheng, Y.-F. The comparison between adenocarcinoma and squamous cell carcinoma in lung cancer patients. *J. Cancer Res. Clin. Oncol.* **2019**, *146*, 43–52. [CrossRef] [PubMed]
6. Travis, W.D. Lung Cancer Pathology. *Clin. Chest Med.* **2020**, *41*, 67–85. [CrossRef] [PubMed]
7. Zhang, J.; Fujimoto, J.; Wedge, D.C.; Song, X.; Seth, S.; Chow, C.-W.; Cao, Y.; Gumbs, C.; Gold, K.A.; Kalhor, N.; et al. Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science* **2014**, *346*, 256–259. [CrossRef] [PubMed]
8. The Cancer Genome Atlas Research Network; Weinstein, J.N.; Collisson, E.A.; Mills, G.B.; Shaw, K.R.M.; Ozenberger, B.A.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J.M. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **2013**, *45*, 1113–1120. [CrossRef] [PubMed]
9. The Cancer Genome Atlas Research Network Comprehensive molecular profiling of lung adenocarcinoma. *Nat. Cell Biol.* **2014**, *511*, 543–550. [CrossRef]
10. The Cancer Genome Atlas Research Network Comprehensive genomic characterization of squamous cell lung cancers. *Nat. Cell Biol.* **2012**, *489*, 519–525. [CrossRef]
11. Krzystanek, M.; Moldvay, J.; Szüts, D.; Szallasi, Z.; Eklund, A.C. A robust prognostic gene expression signature for early stage lung adenocarcinoma. *Biomark. Res.* **2016**, *4*, 1–7. [CrossRef] [PubMed]
12. Shukla, S.; Evans, J.R.; Malik, R.; Feng, F.Y.; Dhanasekaran, S.M.; Cao, X.; Chen, G.; Beer, D.G.; Jiang, H.; Chinnaiyan, A.M. Development of a RNA-Seq Based Prognostic Signature in Lung Adenocarcinoma. *J. Natl. Cancer Inst.* **2017**, *109*, 200. [CrossRef] [PubMed]
13. Li, Z.; Qi, F.; Li, F. Establishment of a Gene Signature to Predict Prognosis for Patients with Lung Adenocarcinoma. *Int. J. Mol. Sci.* **2020**, *21*, 8479. [CrossRef] [PubMed]
14. Zhu, C.-Q.; Strumpf, D.; Li, C.-Y.; Li, Q.; Liu, N.; Der, S.; Shepherd, F.A.; Tsao, M.-S.; Jurisica, I. Prognostic Gene Expression Signature for Squamous Cell Carcinoma of Lung. *Clin. Cancer Res.* **2010**, *16*, 5038–5047. [CrossRef] [PubMed]
15. Li, J.; Wang, J.; Chen, Y.; Yang, L.; Chen, S. A prognostic 4-gene expression signature for squamous cell lung carcinoma. *J. Cell. Physiol.* **2017**, *232*, 3702–3713. [CrossRef] [PubMed]
16. Lu, C.; Chen, H.; Shan, Z.; Yang, L. Identification of differentially expressed genes between lung adenocarcinoma and lung squamous cell carcinoma by gene expression profiling. *Mol. Med. Rep.* **2016**, *14*, 1483–1490. [CrossRef]
17. Wu, X.; Wang, L.; Feng, F.; Tian, S. Weighted gene expression profiles identify diagnostic and prognostic genes for lung adenocarcinoma and squamous cell carcinoma. *J. Int. Med Res.* **2020**, *48*, 0300060519893837. [CrossRef]
18. Colaprico, A.; Silva, T.C.; Olsen, C.; Garofano, L.; Cava, C.; Garolini, D.; Sabedot, T.S.; Malta, T.M.; Pagnotta, S.M.; Castiglioni, I.; et al. TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* **2016**, *44*, e71. [CrossRef] [PubMed]

19. Therneau, T. *A Package for Survival Analysis in R. R Package Version 3.2-7*. 2020. Available online: https://cran.r-project.org/package=survival (accessed on 21 May 2020).
20. Simon, N.; Friedman, J.H.; Hastie, T.; Tibshirani, R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J. Stat. Softw.* **2011**, *39*, 1–13. [CrossRef]
21. Gerds, T.A.; Ozenne, B. *RiskRegression: Risk Regression Models and Prediction Scores for Survival Analysis with Competing Risks. R Package Version 2020.12.08*. 2020. Available online: https://cran.r-project.org/package=riskRegression (accessed on 21 May 2020).
22. Zhang, J.; Jin, Z. *Cutoff: Seek the Significant Cutoff Value. R Package Version 1.3*. 2019. Available online: https://cran.r-project.org/package=cutoff (accessed on 21 May 2020).
23. Ritchie, M.E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C.W.; Shi, W.; Smyth, G.K. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **2015**, *43*, e47. [CrossRef] [PubMed]
24. McCarthy, D.J.; Chen, Y.; Smyth, G.K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **2012**, *40*, 4288–4297. [CrossRef]
25. Vrahatis, A.G.; Balomenos, P.; Tsakalidis, A.K.; Bezerianos, A. DEsubs: An R package for flexible identification of differentially expressed subpathways using RNA-seq experiments. *Bioinformatics* **2016**, *32*, 3844–3846. [CrossRef]
26. Morganella, S.; Pagnotta, S.M.; Ceccarelli, M. *GAIA: An R Package for Genomic Analysis of Significant Chromosomal Aberrations. R Package Version 2.32.0*. 2020. Available online: https://bioconductor.org/packages/gaia (accessed on 21 May 2020).
27. Lawrence, M.; Huber, W.; Pagès, H.; Aboyoun, P.; Carlson, M.; Gentleman, R.; Morgan, M.T.; Carey, V.J. Software for Computing and Annotating Genomic Ranges. *PLoS Comput. Biol.* **2013**, *9*, e1003118. [CrossRef]
28. Silva, T.C.; Colaprico, A.; Olsen, C.; D'Angelo, F.; Bontempi, G.; Ceccarelli, M.; Noushmehr, H. TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages. *F1000Research* **2016**, *5*, 1542. [CrossRef]
29. Mayakonda, A.; Lin, D.-C.; Assenov, Y.; Plass, C.; Koeffler, H.P. Maftools: Efficient and comprehensive analysis of somatic variants in cancer. *Genome Res.* **2018**, *28*, 1747–1756. [CrossRef] [PubMed]
30. Eynden, J.V.D.; Fierro, A.C.; Verbeke, L.P.C.; Marchal, K. SomInaClust: Detection of cancer genes based on somatic mutation patterns of inactivation and clustering. *BMC Bioinform.* **2015**, *16*, 1–12. [CrossRef] [PubMed]
31. Tate, J.G.; Bamford, S.; Jubb, H.C.; Sondka, Z.; Beare, D.M.; Bindal, N.; Boutselakis, H.; Cole, C.G.; Creatore, C.; Dawson, E.; et al. COSMIC: The Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **2018**, *47*, D941–D947. [CrossRef]
32. Zhang, J.; Jin, Z. *Ggrisk: Risk Score Plot for Cox Regression. R Package Version 1.2*. 2020. Available online: https://cran.r-project.org/package=ggrisk (accessed on 21 May 2020).
33. Kassambara, A.; Kosinski, M.; Biecek, P. *Survminer: Drawing Survival Curves Using "ggplot2". R Package Version 0.4.8*. 2020. Available online: https://cran.r-project.org/package=survminer (accessed on 21 May 2020).
34. Heagerty, P.J.; Saha-Chaudhuri, P. *survivalROC: Time-Dependent ROC Curve Estimation from Censored Survival Data. R Package Version 1.0.3*. 2013. Available online: https://cran.r-project.org/package=survivalROC (accessed on 21 May 2020).
35. Kennedy, N. *Forestmodel: Forest Plots from Regression Models. R Package Version 0.6.2*. 2020. Available online: https://cran.r-project.org/package=forestmodel (accessed on 21 May 2020).
36. Durinck, S.; Spellman, P.T.; Birney, E.; Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **2009**, *4*, 1184–1191. [CrossRef]
37. Yu, G.; Wang, L.-G.; Han, Y.; He, Q.-Y. clusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters. *OMICS J. Integr. Biol.* **2012**, *16*, 284–287. [CrossRef]
38. Yu, G. *Enrichplot: Visualization of Functional Enrichment Result. R Package Version 1.8.1*. 2020. Available online: https://github.com/GuangchuangYu/enrichplot (accessed on 21 May 2020).
39. Gu, Z.; Eils, R.; Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **2016**, *32*, 2847–2849. [CrossRef] [PubMed]
40. Meyer, D.; Zeileis, A.; Hornik, K. *Vcd: Visualizing Categorical Data. R Package Version 1.4-8*. 2020. Available online: https://cran.r-project.org/package=vcd (accessed on 21 May 2020).
41. Meyer, D.; Zeileis, A.; Hornik, K. The Strucplot Framework: Visualizing Multi-way Contingency Tables withvcd. *J. Stat. Softw.* **2006**, *17*, 1–48. [CrossRef]
42. Gu, Z.; Gu, L.; Eils, R.; Schlesner, M.; Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* **2014**, *30*, 2811–2812. [CrossRef] [PubMed]
43. Chen, H. *VennDiagram: Generate High-Resolution Venn and Euler Plots. R Package Version 1.6.20*. 2018. Available online: https://cran.r-project.org/package=VennDiagram (accessed on 21 May 2020).
44. Kumar, S.; Rao, N.; Ge, R. Emerging Roles of ADAMTSs in Angiogenesis and Cancer. *Cancers* **2012**, *4*, 1252. [CrossRef]
45. Li, Z.; Weng, H.; Su, R.; Weng, X.; Zuo, Z.; Li, C.; Huang, H.; Nachtergaele, S.; Dong, L.; Hu, C.; et al. FTO Plays an Oncogenic Role in Acute Myeloid Leukemia as a N 6 -Methyladenosine RNA Demethylase. *Cancer Cell* **2017**, *31*, 127–141. [CrossRef]
46. Li, X.; Hu, Z.; Qu, X.; Zhu, J.; Li, L.; Ring, B.Z.; Su, L. Putative EPHX1 Enzyme Activity Is Related with Risk of Lung and Upper Aerodigestive Tract Cancers: A Comprehensive Meta-Analysis. *PLoS ONE* **2011**, *6*, e14749. [CrossRef]
47. Zhu, X.; Guo, X.; Wu, S.; Wei, L. ANGPTL4 Correlates with NSCLC Progression and Regulates Epithelial-Mesenchymal Transition via ERK Pathway. *Lung* **2016**, *194*, 637–646. [CrossRef] [PubMed]
48. Hu, X.-G.; Chen, L.; Wang, Q.-L.; Zhao, X.-L.; Tan, J.; Cui, Y.-H.; Liu, X.-D.; Zhang, X.; Bian, X.-W. Elevated expression of ASCL2 is an independent prognostic indicator in lung squamous cell carcinoma. *J. Clin. Pathol.* **2015**, *69*, 313–318. [CrossRef]

49. Kadomoto, S.; Izumi, K.; Mizokami, A. The CCL20-CCR6 Axis in Cancer Progression. *Int. J. Mol. Sci.* **2020**, *21*, 5186. [CrossRef]
50. Zhang, J.; Zhang, X.; Zhao, X.; Jiang, M.; Gu, M.; Wang, Z.; Yue, W. DKK1 promotes migration and invasion of non–small cell lung cancer via β-catenin signaling pathway. *Tumor Biol.* **2017**, *39*. [CrossRef] [PubMed]
51. Inoue, R.; Hirohashi, Y.; Kitamura, H.; Nishida, S.; Murai, A.; Takaya, A.; Yamamoto, E.; Matsuki, M.; Tanaka, T.; Kubo, T.; et al. GRIK2 has a role in the maintenance of urothelial carcinoma stem-like cells, and its expression is associated with poorer prognosis. *Oncotarget* **2017**, *8*, 28826–28839. [CrossRef] [PubMed]
52. Yu, C.; Hou, L.; Cui, H.; Zhang, L.; Tan, X.; Leng, X.; Li, Y. LDHA upregulation independently predicts poor survival in lung adenocarcinoma, but not in lung squamous cell carcinoma. *Futur. Oncol.* **2018**, *14*, 2483–2492. [CrossRef]
53. Yang, L.; Lee, M.M.; Leung, M.M.; Wong, Y.H. Regulator of G protein signaling 20 enhances cancer cell aggregation, migration, invasion and adhesion. *Cell. Signal.* **2016**, *28*, 1663–1672. [CrossRef]
54. Han, S.-W.; Kim, H.-P.; Shin, J.-Y.; Jeong, E.-G.; Lee, W.-C.; Kim, K.Y.; Park, S.Y.; Lee, D.-W.; Won, J.-K.; Jeong, S.-Y.; et al. RNA editing in RHOQ promotes invasion potential in colorectal cancer. *J. Exp. Med.* **2014**, *211*, 613–621. [CrossRef]
55. Yuan, D.; Yang, X.; Yuan, Z.; Zhao, Y.; Guo, J. TLE1 function and therapeutic potential in cancer. *Oncotarget* **2016**, *8*, 15971–15976. [CrossRef]
56. Tabatabaeian, H.; Rao, A.; Ramos, A.; Chu, T.; Sudol, M.; Lim, Y.P. The emerging roles of WBP2 oncogene in human cancers. *Oncogene* **2020**, *39*, 4621–4635. [CrossRef] [PubMed]
57. Yoshimura, K.; Suzuki, Y.; Inoue, Y.; Tsuchiya, K.; Karayama, M.; Iwashita, Y.; Kahyo, T.; Kawase, A.; Tanahashi, M.; Ogawa, H.; et al. CD200 and CD200R1 are differentially expressed and have differential prognostic roles in non-small cell lung cancer. *OncoImmunology* **2020**, *9*, 1746554. [CrossRef]
58. Gao, C.; Zhuang, J.; Li, H.; Liu, C.; Zhou, C.; Liu, L.; Sun, C. Exploration of methylation-driven genes for monitoring and prognosis of patients with lung adenocarcinoma. *Cancer Cell Int.* **2018**, *18*, 1–11. [CrossRef] [PubMed]
59. Zheng, X.; Li, Y.; Ma, C.; Zhang, J.; Zhang, Y.; Fu, Z.; Luo, H. Independent Prognostic Potential of GNPNAT1 in Lung Adenocarcinoma. *BioMed Res. Int.* **2020**, *2020*, 1–16. [CrossRef] [PubMed]
60. Wang, Q.; Qiu, X. Comprehensive Analysis of the Expression and Prognosis for IRXs in Non-small Cell Lung Cancer. *Res. Sq.* **2020**. [CrossRef]
61. Puderecki, M.; Szumiło, J.; Marzec-Kotarska, B. Novel prognostic molecular markers in lung cancer (Review). *Oncol. Lett.* **2020**, *20*, 9–18. [CrossRef]
62. Zhao, R.; Ding, D.; Yu, W.; Zhu, C.; Ding, Y. The Lung Adenocarcinoma Microenvironment Mining and Its Prognostic Merit. *Technol. Cancer Res. Treat.* **2020**, *19*. [CrossRef]
63. Wang, L.; Zhao, H.; Xu, Y.; Li, J.; Deng, C.; Deng, Y.; Bai, J.; Li, X.; Xiao, Y.; Zhang, Y. Systematic identification of lincRNA-based prognostic biomarkers by integrating lincRNA expression and copy number variation in lung adenocarcinoma. *Int. J. Cancer* **2019**, *144*, 1723–1734. [CrossRef] [PubMed]
64. Wright, C.M.; Francis, S.M.S.; Tan, M.E.; Martins, M.U.; Winterford, C.; Davidson, M.R.; Duhig, E.E.; Clarke, B.E.; Hayward, N.K.; Yang, I.A.; et al. MS4A1 Dysregulation in Asbestos-Related Lung Squamous Cell Carcinoma Is Due to CD20 Stromal Lymphocyte Expression. *PLoS ONE* **2012**, *7*, e34943. [CrossRef] [PubMed]
65. Qi, J.; Mu, D. MicroRNAs and lung cancers: From pathogenesis to clinical implications. *Front. Med.* **2012**, *6*, 134–155. [CrossRef]
66. Sage, A.P.; Ng, K.W.; Marshall, E.A.; Stewart, G.L.; Minatel, B.C.; Enfield, K.S.S.; Martin, S.D.; Brown, C.J.; Abraham, N.; Lam, W.L. Assessment of long non-coding RNA expression reveals novel mediators of the lung tumour immune response. *Sci. Rep.* **2020**, *10*, 1–13. [CrossRef]
67. Tang, L.-X.; Chen, G.-H.; Li, H.; He, P.; Zhang, Y.; Xu, X.-W. Long non-coding RNA OGFRP1 regulates LYPD3 expression by sponging miR-124-3p and promotes non-small cell lung cancer progression. *Biochem. Biophys. Res. Commun.* **2018**, *505*, 578–585. [CrossRef] [PubMed]
68. Giacalone, N.J.; Den, R.B.; Eisenberg, R.; Chen, H.; Olson, S.J.; Massion, P.P.; Carbone, D.P.; Lu, B. ALDH7A1 expression is associated with recurrence in patients with surgically resected non-small-cell lung carcinoma. *Futur. Oncol.* **2013**, *9*, 737–745. [CrossRef] [PubMed]
69. Wang, J.; Shen, Q.; Shi, Q.; Yu, B.; Wang, X.; Cheng, K.; Lu, G.; Zhou, X. Detection of ALK protein expression in lung squamous cell carcinomas by immunohistochemistry. *J. Exp. Clin. Cancer Res.* **2014**, *33*, 1–7. [CrossRef] [PubMed]
70. Boldrini, L.; Gisfredi, S.; Ursino, S.; Faviana, P.; Lucchi, M.; Melfi, F.; Mussi, A.; Basolo, F.; Fontanini, G. Expression of endothelin-1 is related to poor prognosis in non-small cell lung carcinoma. *Eur. J. Cancer* **2005**, *41*, 2828–2835. [CrossRef] [PubMed]
71. Zhang, Y.; Zhao, X.; Deng, L.; Li, X.; Wang, G.; Li, Y.; Chen, M. High expression of FABP4 and FABP6 in patients with colorectal cancer. *World J. Surg. Oncol.* **2019**, *17*, 1–13. [CrossRef]
72. Wang, X.; Shi, B.; Zhao, Y.; Lu, Q.; Fei, X.; Lu, C.; Li, C.; Chen, H. HKDC1 promotes the tumorigenesis and glycolysis in lung adenocarcinoma via regulating AMPK/mTOR signaling pathway. *Cancer Cell Int.* **2020**, *20*, 1–12. [CrossRef]
73. Guan, Y.; Wang, Y.; Bhandari, A.; Xia, E.; Wang, O. IGSF1: A novel oncogene regulates the thyroid cancer progression. *Cell Biochem. Funct.* **2019**, *37*, 516–524. [CrossRef]
74. Gong, J.; Tian, J.; Lou, J.; Wang, X.; Ke, J.; Li, J.; Yang, Y.; Gong, Y.; Zhu, Y.; Zou, D.; et al. A polymorphic MYC response element in KBTBD11 influences colorectal cancer risk, especially in interaction with an MYC-regulated SNP rs6983267. *Ann. Oncol.* **2017**, *29*, 632–639. [CrossRef]

75. Zou, Z.; Li, X.; Sun, Y.; Li, L.; Zhang, Q.; Zhu, L.; Zhong, Z.; Wang, M.; Wang, Q.; Liu, Z.; et al. NOS1 expression promotes proliferation and invasion and enhances chemoresistance in ovarian cancer. *Oncol. Lett.* **2020**, *19*, 2989–2995. [CrossRef]
76. Ueda, M.; Iguchi, T.; Masuda, T.; Komatsu, H.; Nambara, S.; Sakimura, S.; Hirata, H.; Uchi, R.; Eguchi, H.; Ito, S.; et al. Up-regulation of SLC9A9 Promotes Cancer Progression and Is Involved in Poor Prognosis in Colorectal Cancer. *Anticancer Res.* **2017**, *37*, 2255–2263. [CrossRef]
77. Huang, N.; Lin, W.; Shi, X.; Tao, T. STK24 expression is modulated by DNA copy number/methylation in lung adenocarcinoma and predicts poor survival. *Futur. Oncol.* **2018**, *14*, 2253–2263. [CrossRef]
78. Tang, Y.; Geng, Y.; Luo, J.; Shen, W.; Zhu, W.; Meng, C.; Li, M.; Zhou, X.; Zhang, S.; Cao, J. Downregulation of ubiquitin inhibits the proliferation and radioresistance of non-small cell lung cancer cells in vitro and in vivo. *Sci. Rep.* **2015**, *5*, 1–12. [CrossRef]
79. Baykara, O.; Dalay, N.; Kaynak, K.; Buyru, N. ZNF703 Overexpression may act as an oncogene in non-small cell lung cancer. *Cancer Med.* **2016**, *5*, 2873–2878. [CrossRef]
80. Li, J.; Ye, L.; Mansel, R.E.; Jiang, W.G. Potential prognostic value of repulsive guidance molecules in breast cancer. *Anticancer Res.* **2011**, *31*, 1703–1711. [PubMed]
81. Chong, P.K.; Lee, H.; Zhou, J.; Liu, S.-C.; Loh, M.C.S.; Wang, T.T.; Chan, S.P.; Smoot, D.T.; Ashktorab, H.; So, J.B.Y.; et al. ITIH3 Is a Potential Biomarker for Early Detection of Gastric Cancer. *J. Proteome Res.* **2010**, *9*, 3671–3679. [CrossRef] [PubMed]
82. Liu, Y.; Cui, J.; Tang, Y.-L.; Huang, L.; Zhou, C.-Y.; Xu, J.-X. Prognostic Roles of mRNA Expression of S100 in Non-Small-Cell Lung Cancer. *BioMed Res. Int.* **2018**, *2018*, 1–11. [CrossRef] [PubMed]
83. Jia, Z.; Gao, S.; M'Rabet, N.; De Geyter, C.; Zhang, H. Sp1 Is Necessary for Gene Activation of Adamts17 by Estrogen. *J. Cell. Biochem.* **2014**, *115*, 1829–1839. [CrossRef]
84. Li, R.; Yang, Y.-E.; Jin, J.; Zhang, M.-Y.; Liu, X.-X.; Yin, Y.-H.; Qu, Y.-Q. Identification of lncRNA biomarkers in lung squamous cell carcinoma using comprehensive analysis of lncRNA mediated ceRNA network. *Artif. Cells Nanomed. Biotechnol.* **2019**, *47*, 3246–3258. [CrossRef] [PubMed]
85. Han, B.-W.; Ye, H.; Wei, P.-P.; He, B.; Han, C.; Chen, Z.-H.; Chen, Y.-Q.; Wang, W.-T. Global identification and characterization of lncRNAs that control inflammation in malignant cholangiocytes. *BMC Genom.* **2018**, *19*, 1–13. [CrossRef] [PubMed]
86. Rao, Y.; Liu, H.; Yan, X.; Wang, J. In Silico Analysis Identifies Differently Expressed lncRNAs as Novel Biomarkers for the Prognosis of Thyroid Cancer. *Comput. Math. Methods Med.* **2020**, *2020*, 1–10. [CrossRef] [PubMed]
87. Zhang, W.; Ye, Y.J.; Ren, X.W.; Huang, J.; Shen, Z.L. Detection of preoperative chemoradiotherapy sensitivity molecular characteristics of rectal cancer by transcriptome second generation sequencing. *J. Peking Univ. Health Sci.* **2019**, *51*, 542–547. [CrossRef]
88. Tian, B.; Han, X.; Li, G.; Jiang, H.; Qi, J.; Li, J.; Tian, Y.; Wang, C. A Long Intergenic Non-coding RNA, LINC01426, Promotes Cancer Progression via AZGP1 and Predicts Poor Prognosis in Patients with LUAD. *Mol. Ther. Methods Clin. Dev.* **2020**, *18*, 765–780. [CrossRef]
89. Yang, H.; Jin, L.; Sun, X. A thirteen-gene set efficiently predicts the prognosis of glioblastoma. *Mol. Med. Rep.* **2019**, *19*, 1613–1621. [CrossRef]
90. Hanahan, D.; Weinberg, R.A. Hallmarks of Cancer: The Next Generation. *Cell* **2011**, *144*, 646–674. [CrossRef] [PubMed]
91. Phan, L.M.; Yeung, S.J.; Lee, M.-H. Cancer metabolic reprogramming: Importance, main features, and potentials for precise targeted anti-cancer therapies. *Cancer Biol. Med.* **2014**, *11*, 1–19.
92. Keenan, M.M.; Chi, J.-T. Alternative Fuels for Cancer Cells. *Cancer J.* **2015**, *21*, 49–55. [CrossRef] [PubMed]
93. Hamada, K.; Kohno, T.; Kawanishi, M.; Ohwada, S.; Yokota, J. Association ofCDKN2A (p16)/CDKN2B (p15) alterations and homozygous chromosome arm 9p deletions in human lung carcinoma. Genes, Chromosom. *Cancer* **1998**, *22*, 232–240. [CrossRef]
94. Schuster, K.; Venkateswaran, N.; Rabellino, A.; Girard, L.; Peña-Llopis, S.; Scaglioni, P.P. Nullifying the CDKN2AB Locus Promotes Mutant K-ras Lung Tumorigenesis. *Mol. Cancer Res.* **2014**, *12*, 912–923. [CrossRef]
95. Jiang, J.; Gu, Y.; Liu, J.; Wu, R.; Fu, L.; Zhao, J.; Guan, Y. Coexistence of p16/CDKN2A homozygous deletions and activating EGFR mutations in lung adenocarcinoma patients signifies a poor response to EGFR-TKIs. *Lung Cancer* **2016**, *102*, 101–107. [CrossRef] [PubMed]
96. Bass, A.J.; Watanabe, H.; Mermel, C.H.; Yu, S.; Perner, S.; Verhaak, R.G.; Kim, S.Y.; Wardwell, L.; Tamayo, P.; Gat-Viks, I.; et al. SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas. *Nat. Genet.* **2009**, *41*, 1238–1242. [CrossRef]
97. Maier, S.; Wilbertz, T.; Braun, M.; Scheble, V.; Reischl, M.; Mikut, R.; Menon, R.; Nikolov, P.; Petersen, K.; Beschorner, C.; et al. SOX2 amplification is a common event in squamous cell carcinomas of different organ sites. *Hum. Pathol.* **2011**, *42*, 1078–1088. [CrossRef]
98. Fejzo, M.S.; Anderson, L.; Chen, H.-W.; Guandique, E.; Kalous, O.; Conklin, D.; Slamon, D.J. Proteasome ubiquitin receptor PSMD4 is an amplification target in breast cancer and may predict sensitivity to PARPi. Genes, Chromosom. *Cancer* **2017**, *56*, 589–597. [CrossRef]
99. Cai, M.-J.; Cui, Y.; Fang, M.; Wang, Q.; Zhang, A.-J.; Kuai, J.-H.; Pang, F.; Cui, X.-D. Inhibition of PSMD4 blocks the tumorigenesis of hepatocellular carcinoma. *Gene* **2019**, *702*, 66–74. [CrossRef]
100. Cheng, Y.-M.; Lin, P.-L.; Wu, D.-W.; Wang, L.; Huang, C.-C.; Lee, H. PSMD4 is a novel therapeutic target in chemoresistant colorectal cancer activated by cytoplasmic localization of Nrf2. *Oncotarget* **2018**, *9*, 26342–26352. [CrossRef] [PubMed]

101. Türkoğlu, S.A.; Dayi, G.; Köçkar, F. Upregulation of PSMD4 Gene By Hypoxia in Prostate Cancer Cells. *Turk. J. Boil.* **2020**, *44*, 275–283. [CrossRef]
102. O'Leary, C.; Gasper, H.; Sahin, K.B.; Tang, M.; Kulasinghe, A.; Adams, M.N.; Richard, D.J.; O'Byrne, K.J. Epidermal Growth Factor Receptor (EGFR)-Mutated Non-Small-Cell Lung Cancer (NSCLC). *Pharmaceuticals* **2020**, *13*, 273. [CrossRef]
103. Mathsyaraja, H.; Catchpole, J.; Eastwood, E.; Babaeva, E.; Geuenich, M.; Cheng, P.F.; Freie, B.; Ayers, J.; Yu, M.; Wu, N.; et al. Loss of MGA mediated Polycomb repression promotes tumor progression and invasiveness. *bioRxiv* **2020**. [CrossRef]
104. Xue, Y.; Meehan, B.; Fu, Z.; Wang, X.Q.D.; Fiset, P.O.; Rieker, R.; Levins, C.; Kong, T.; Zhu, X.; Morin, G.; et al. SMARCA4 loss is synthetic lethal with CDK4/6 inhibition in non-small cell lung cancer. *Nat. Commun.* **2019**, *10*, 1–13. [CrossRef]
105. Xu, Y.; Gao, P.; Lv, X.; Zhang, L.; Zhang, J. The role of the ataxia telangiectasia mutated gene in lung cancer: Recent advances in research. *Ther. Adv. Respir. Dis.* **2017**, *11*, 375–380. [CrossRef] [PubMed]
106. Sun, X.; Jia, M.; Sun, W.; Feng, L.; Gu, C.; Wu, T. Functional role of RBM10 in lung adenocarcinoma proliferation. *Int. J. Oncol.* **2018**, *54*, 467–478. [CrossRef]
107. Chang, S.; Yim, S.; Park, H. The cancer driver genes IDH1/2, JARID1C/ KDM5C, and UTX/ KDM6A: Crosstalk between histone demethylation and hypoxic reprogramming in cancer metabolism. *Exp. Mol. Med.* **2019**, *51*, 1–17. [CrossRef] [PubMed]
108. Tam, K.W.; Zhang, W.; Soh, J.; Stastny, V.; Chen, M.; Sun, H.; Thu, K.; Rios, J.J.; Yang, C.; Marconett, C.N.; et al. CDKN2A/p16 Inactivation Mechanisms and Their Relationship to Smoke Exposure and Molecular Features in Non–Small-Cell Lung Cancer. *J. Thorac. Oncol.* **2013**, *8*, 1378–1388. [CrossRef]
109. Gkountakos, A.; Sartori, G.; Falcone, I.; Piro, G.; Ciuffreda, L.; Carbone, C.; Tortora, G.; Scarpa, A.; Bria, E.; Milella, M.; et al. PTEN in Lung Cancer: Dealing with the Problem, Building on New Knowledge and Turning the Game Around. *Cancers* **2019**, *11*, 1141. [CrossRef] [PubMed]
110. Pązik, M.; Michalska, K.; Żebrowska-Nawrocka, M.; Zawadzka, I.; Łochowski, M.; Balcerczak, E. Clinical significance of HRAS and KRAS genes expression in patients with non–small-cell lung cancer—Preliminary Findings. *BMC Cancer* **2021**, *21*, 1–13. [CrossRef] [PubMed]
111. Zhao, J.; Han, Y.; Li, J.; Chai, R.; Bai, C. Prognostic value of KRAS/TP53/PIK3CA in non-small cell lung cancer. *Oncol. Lett.* **2019**, *17*, 3233–3240. [CrossRef] [PubMed]
112. Liu, P.; Morrison, C.; Wang, L.; Xiong, D.; Vedell, P.; Cui, P.; Hua, X.; Ding, F.; Lu, Y.; James, M.; et al. Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing. *Carcinogenesis* **2012**, *33*, 1270–1276. [CrossRef] [PubMed]
113. La Fleur, L.; Falk-Sörqvist, E.; Smeds, P.; Berglund, A.; Sundström, M.; Mattsson, J.S.; Brandén, E.; Koyi, H.; Isaksson, J.; Brunnström, H.; et al. Mutation patterns in a population-based non-small cell lung cancer cohort and prognostic impact of concomitant mutations in KRAS and TP53 or STK11. *Lung Cancer* **2019**, *130*, 50–58. [CrossRef] [PubMed]

APPENDIX B.

ARTICLE II

TCGAnalyzeR: a web application for integrative visualization of molecular and clinical data of cancer patients for cohort and associated gene discovery

Talip Zengin, Başak Abak Masud, Tuğba Önal-Süzek

bioRxiv, 2023, doi: 10.1101/2023.01.20.524925

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.20.524925; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

TCGAnalyzeR: a web application for integrative visualization of molecular and clinical data of cancer patients for cohort and associated gene discovery

Talip Zengin[1,2,☯], Başak Abak Masud[2,☯], Tuğba Önal-Süzek[2,3,*]

[1]Department of Molecular Biology and Genetics, Mugla Sitki Kocman University, Mugla, 48000, Turkiye [2]Department of Bioinformatics, Mugla Sitki Kocman University, Mugla, 48000, Turkiye, [3]Department of Computer Engineering, Mugla Sitki Kocman University, Mugla, 48000, Turkiye.

*Corresponding Author.

☯Joint first authors.

Abstract

Motivation: The vast size and complexity of The Cancer Genome Atlas (TCGA) database with multidimensional molecular and clinical data of ~11,000 cancer patients of 33 cancer types challenge the effective utilization of this valuable resource. Therefore, we built a web application named TCGAnalyzeR with the main idea of presenting an integrative visualization of mutations, transcriptome profile, copy number variation and clinical data allowing researchers to facilitate the identification of customized patient cohorts and gene sets for better decision-making for oncologists and cancer researchers.

Results: We present TCGAnalyzeR for integrative visualization of pre-analyzed TCGA data with the several novel modules: (i) Simple nucleotide variations with driver prediction; (ii) Recurrent copy number alterations; (iii) Differential expression in tumor versus normal, with pathway enrichment and the survival analysis; (iii) TCGA clinical data and survival analysis; (iv) External subcohorts from literature, curatedTCGAData and BiocOncoTK R packages; (v) Internal patient clusters determined using iClusterPlus R package or signature-based expression analysis. TCGAnalyzeR provides clinical oncologists and cancer researchers interactive and integrative representations of these multi-omic, pan-cancer TCGA data with availability of subcohort analysis and visualization. TCGAnalyzeR can be used to create their own custom gene sets for pan-cancer comparisons, to create custom patient subcohorts comparing external subcohorts (MSI, Immune, PAM50, Triple Negative, IDH1, miRNA, etc) along with our internal patient clusters, to visualize cohort-centric or gene-centric results along with pathway enrichment and survival analysis graphically on an interactive web tool.

Availability: TCGAnalyzeR is freely available on the web at http://tcganalyzer.mu.edu.tr.

Contact: tugbasuzek@mu.edu.tr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The sheer scale and complexity of The Cancer Genome Atlas (TCGA) data (The Cancer Genome Atlas Research Network. *et al.*, 2013) offers great potential for scientific discovery, but the challenges to effective use of this valuable resource by biologists and clinicians have led to the development of several visualization tools such as cBioPortal (Cerami *et al.*, 2012; Gao *et al.*, 2013), Firebrowse (Deng *et al.*, 2017), and University of California, Santa Cruz (UCSC) Xena (Goldman *et al.*, 2020). Among these tools, cBioPortal is the most preferred due to its interactive exploration of larger and up-to-date cancer datasets. OncoKB (Chakravarty et al., 2017) is another precision oncology knowledge base that allows searching and comparing drug response data from different TCGA cohorts. Although the ICGC web portal (Zhang et al., 2019) and Coral web application (Adelberger *et al.*, 2021) allow patient/gene subsetting of TCGA cohorts and provide survival and Venn diagram visualization of cohorts, they do not allow comparison of cohorts pre-generated by other research groups. These tools address the reuse needs of users. However, they only provide raw data without allowing the patient/gene subsetting of the statistical analyses for pan-cancer subcohort and associated gene discovery.

We built an interactive Shiny (Chang *et al.*, 2022) web application for the analysis and visualization of four data categories across 33 cancer types. Users can visualize the results of preprocessed analysis of Simple Nucleotide Variations (SNVs), Copy Number Variations (CNVs), differential gene expression in tumor versus normal samples, and clinical data of TCGA projects from National Cancer Institute (NCI) Genomic Data Commons (GDC) (Grossman *et al.*, 2016). Moreover, users can compare patient clusters determined using iClusterPlus R package (Mo and Shen, 2022) with expression-based survival risk groups (Zengin and Önal-Süzek, 2020, 2021), and curated subtypes such as immune subtypes (Thorsson *et al,* 2019), Triple Negative Breast Cancer (TNBC)

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.20.524925; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

subtypes (Lehmann *et al*, 2016), PAM50 subtypes (Berger *et al*, 2018) and Microsatellite Instability (MSI) related subgroups and several data type clusters from BiocOncoTK (Carey, 2022; Ding *et al*, 2018) and *curatedTCGAData* (Ramos *et al.*, 2020) R packages. Furthermore, users can create custom subcohorts based on genomic analyses and/or clinical data to subset data visualization. Users can also create gene sets for data type and/or pan-cancer comparisons. For each cancer, whenever available, sample types, survival risk groups (Low-risk / High-risk), and pre-computed or curated patient clusters can be used for filtering patients. The main novelty of our tool is allowing the users to generate custom patient sub-cohorts and/or gene sets using interactive graphical representations via clipboard functionality.

2 Methods

2.1 TCGA data

Publicly available hg38 data including SNV, CNV, Transcriptome Profiling, microRNA, Methylation, and clinical data of 33 cancer types from The Cancer Genome Atlas (TCGA) projects were downloaded on March 6, 2022 from NCI GDC (Grossman *et al.*, 2016) using TCGAbiolinks R package (Colaprico *et al.*, 2016).

2.2 Pre-computed Molecular Data Analysis

2.2.1 SNV Analysis

Potential driver mutated genes with their roles as a tumor suppressor or oncogene were determined by SomInaClust R package (Van den Eynden *et al.*, 2015) using mutation annotation format (maf) file generated by mutect2 pipeline. With the “Somatic Driver Mutations” option, the user can see the significant mutated genes ranked by their Q-value. This option is only available for the “SNV Analysis” category.

2.2.2 CNV Analysis

Significant recurrent copy number variations were identified by *GAIA* R package (Morganella *et al.*, 2011). NCBI IDs and Hugo Symbols of the genes on chromosomal regions with altered copy numbers were determined using *GenomicRanges* (Lawrence *et al.*, 2013) and *biomaRt* (Durinck *et al.*, 2009) R packages.

2.2.3 Differential Gene Expression Analysis

Differentially expressed genes were determined using normalized HTseq counts, by limma-voom method with duplicate-correlation function from *edgeR* (Robinson *et al.*, 2010) and *limma* (Ritchie *et al.*, 2015) R packages. Ensembl IDs were converted to NCBI IDs and Hugo Symbols using the *biomaRt* package (Durinck *et al.*, 2009).

Two different analyses were performed using paired tumor samples against tumor-adjacent normal samples of patients with both sample types (Paired), or tumor samples of all patients against normal samples of patients who have both sample types (All) if it is available for a particular cancer.

2.2.4 Pathway Enrichment

Pathway enrichment and visualization was performed for each analysis by *clusterProfiler* R package (Yu *et al.*, 2012).

2.3 Pre-computed Patient Clusters and Sample Subtypes

TCGAnalyzerR provides an interactive visual analysis of several patient cohorts: i) Survival Risk Groups: We provide low-risk or high-risk patient groups determined by expression-based gene signature analysis for Lung Adenocarcinoma (LUAD), Lung Squamous Cell Carcinoma (LUSC) and Colon Adenocarcinoma (COAD) (Zengin and Önal-Süzek, 2020, 2021), ii) iClusters: We clustered patients using their raw SNV, CNV, gene expression, miRNA expression and methylation data of tumor samples which have all types of data by iClusterBayes method (Mo *et al.*, 2018), iii) Curated subtypes (immune subtypes, TNBC subtypes, PAM50 subtypes) from original publications (Thorsson *et al,* 2019; Lehmann *et al*, 2016; Berger *et al*, 2018). For fifteen cancer types, previously published TCGA cohorts of the individual tumor types are retrieved from curatedTCGA R package curatedTCGAData R package (Ramos *et al.*, 2020). Patient clusters based on Microsatellite Instability (MSI) were compiled using BiocOncoTK (Carey, 2022, Ding *et al.*, 2018) and Immune clusters (Thorsson et al, 2019) were compiled for all 33 cancers. In total, 123 external patient cohorts are integrated into the web interface allowing efficient filtering and cross-comparative analysis of multiple subcohorts in parallel.

2.4 Survival Analysis

Kaplan-Meier (KM) survival analysis is performed by *readr* (Wickham *et al.*, 2022) and *survfit* (Therneau, 2022) R packages in real-time based on overall survival data of patients of interest for selected clinical features.

2.5 Visualization

TCGAnalyzeR front-end was implemented by javascript-based R packages with an interactive dashboard enabling users to select cancer types, data types, risk groups and patient cohorts using *heatmaply, g3viz* and *highcharter* R packages (Galili *et al.*, 2018; Guo *et al.*, 2019; Kunst, 2022). All visualizations are interactive and customizable by the user through the filtration options with “My genes” and/or “My patients” panels enabling to copy genes and/or patients of interest to the clipboard. TCGAnalyzeR currently supports TSV for downloading tables; and high-resolution PNG format for downloading figures.

3 Results

TCGAnalyzeR web application offers simple nucleotide (SNV) analysis as the first step. We present two data sets for SNV analysis: ‘Somatic Driver Mutations’ predicted by the SomInaClust R package and ‘All’ mutations from the original maf file without any analysis. Oncoplot in Figure 1 shows candidate driver genes with their percentage in tumor samples of Breast Invasive Carcinoma (BRCA) with annotations of patient iClusters, PAM50, TNBC and immune subtypes. iCluster #1 is highly correlated with the Basal and TNBC subtype. Wound-healing and IFNɣ-dominant immune subtypes gather around iCluster #1. iCluster #2 is mostly correlated with Luminal A subtype and Inflammatory immune subtype. iCluster #3 seems to be a mixture of estrogen receptor positive Luminal A and Luminal B subtypes and heterogenous immune subtypes. Moreover, both iCluster #1 and iCluster #2 are not TNBC subtypes. On the other hand, iCluster #1 shows a highly different mutation pattern than other clusters. iCluster #1 together with basal and triple-negative subtypes have higher prevalence of TP53 mutations with very few mutations of PIK3CA, CDH1, GATA3, KMT2C, MAP3K1 genes. Moreover, mutations of TP53, CDH1 and GATA3 genes are mutually exclusive (Figure 1).

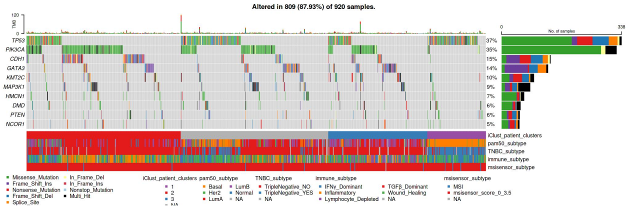
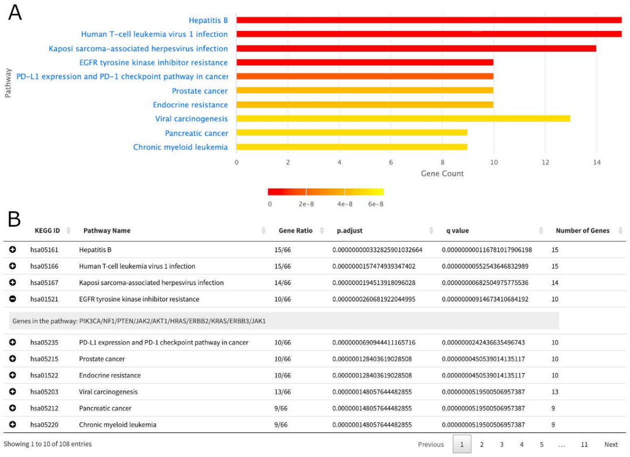


Fig. 1. Oncoplot of candidate driver genes with patient clusters and sample subtypes. Top10 significant candidate driver genes with mutations determined by SomInaClust R package. Bottom annotations show the sample subtypes curated from the literature and iClusters.

Pathway enrichment of candidate driver mutated genes is shown as a bar graph in Figure 2A. Significant pathways of driver genes are highly cancer-related pathways such as EGFR tyrosine kinase inhibitor resistance, PD-L1 and PD-1 pathway in cancer, prostate cancer, pancreatic cancer and chronic myeloid leukemia pathways. Pathway enrichment analysis also supplies a table showing KEGG IDs, with related genes and p/q-values (Figure 2B).

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.20.524925; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



	KEGG ID	Pathway Name	Gene Ratio	p.adjust	q value	Number of Genes
⊕	hsa05161	Hepatitis B	15/66	0.00000000332825901032664	0.00000000116781017906198	15
⊕	hsa05166	Human T-cell leukemia virus 1 infection	15/66	0.0000000157474939347402	0.0000000552543646832989	15
⊕	hsa05167	Kaposi sarcoma-associated herpesvirus infection	14/66	0.0000000194513918096028	0.0000000682504975775536	14
⊖	hsa01521	EGFR tyrosine kinase inhibitor resistance	10/66	0.0000000260681922044995	0.0000000914673410684192	10
	Genes in the pathway: PIK3CA/NF1/PTEN/JAK2/AKT1/HRAS/ERBB2/KRAS/ERBB3/JAK1					
⊕	hsa05235	PD-L1 expression and PD-1 checkpoint pathway in cancer	10/66	0.0000000690944411165716	0.0000000242436635496743	10
⊕	hsa05215	Prostate cancer	10/66	0.000000128403619028508	0.0000000450539014135117	10
⊕	hsa01522	Endocrine resistance	10/66	0.000000128403619028508	0.0000000450539014135117	10
⊕	hsa05203	Viral carcinogenesis	13/66	0.000000148057644482855	0.0000000519500506957387	13
⊕	hsa05212	Pancreatic cancer	9/66	0.000000148057644482855	0.0000000519500506957387	9
⊕	hsa05220	Chronic myeloid leukemia	9/66	0.000000148057644482855	0.0000000519500506957387	9

Showing 1 to 10 of 108 entries

Previous 1 2 3 4 5 … 11 Next

Fig. 2. Pathway enrichment of candidate driver genes. A. Bar plot showing top 10 significant pathways of candidate driver genes determined by SomInaClust R package. B. Pathway enrichment table presenting KEGG ID, genes in significant pathways with adjusted p-value and q-value.

The “My genes” clipboard panel of TCGAnalyzeR allows modifying plots to show genes of interest. For example, Figure 3 shows the mutation pattern of Oncotype DX gene set together with clinical annotations. iCluster #2, Luminal A subtype and Her2 subtypes are related with ERBB2 (HER2) mutations. Besides, iCluster #1 have fewer mutations than the other two iClusters. Moreover, mutations of Oncotype DX genes are mostly mutually exclusive (Figure 3).

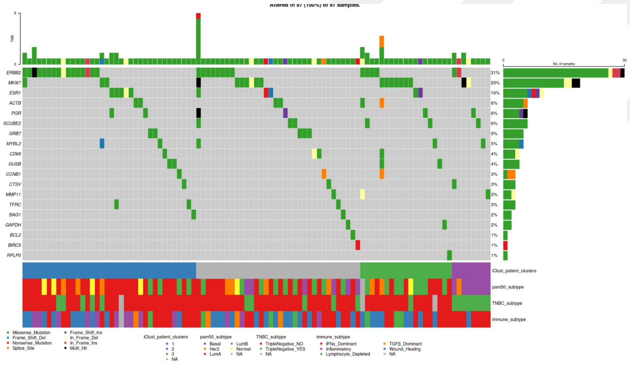
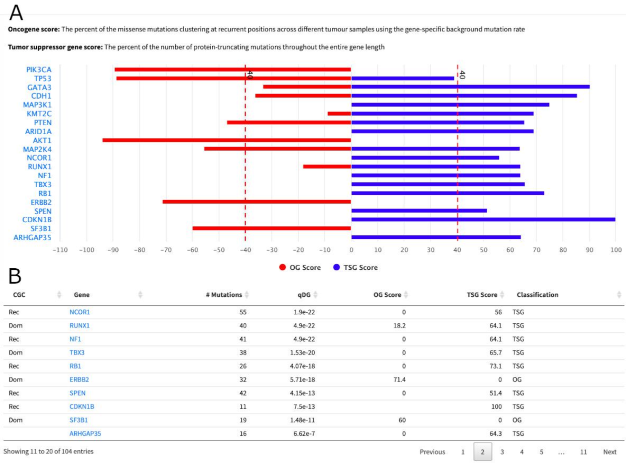


Fig. 3. Oncoplot of Oncotype DX genes with patient clusters and sample subtypes. Mutations of Oncotype DX genes with annotations showing the patient iClusters and sample subtypes curated from the literature.

SomInaClust R package determines candidate driver mutated genes with their potential roles as a tumor suppressor (TSG) or oncogene (OG) with predicted scores (Van den Eynden *et al.*, 2015). Pyramid plot in Figure 4A summarizes OG score and TSG scores of candidate driver genes ranked by their analysis q-values. Some genes may have both OG score and TSG score over threshold, in that case, SomInaClust considers the COSMIC cancer gene census information. Only 1 (ERBB2) of 21 Oncotype DX genes were predicted as significant driver mutated genes. ERBB2 is predicted as an oncogene by SomInaClust as listed Dominant (OG) in COSMIC cancer gene census (Figure 4B).



Oncogene score: The percent of the missense mutations clustering at recurrent positions across different tumour samples using the gene-specific background mutation rate

Tumor suppressor gene score: The percent of the number of protein-truncating mutations throughout the entire gene length

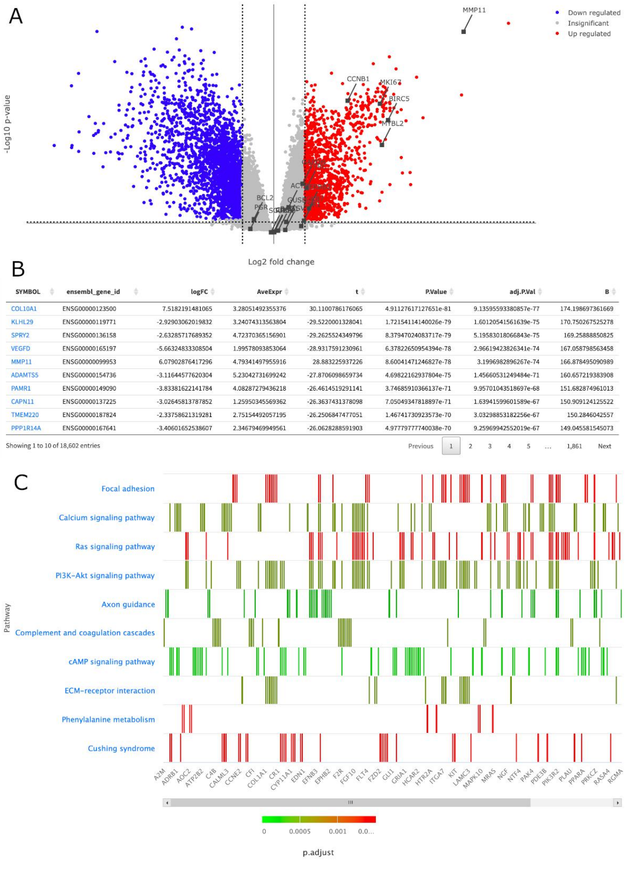
CGC	Gene	# Mutations	qDG	OG Score	TSG Score	Classification
Rec	NCOR1	55	1.9e-22	0	56	TSG
Dom	RUNX1	40	4.9e-22	18.2	64.1	TSG
Rec	NF1	41	4.9e-22	0	64.1	TSG
Dom	TBX3	38	1.53e-20	0	65.7	TSG
Rec	RB1	26	4.07e-18	0	73.1	TSG
Dom	ERBB2	32	5.71e-18	71.4	0	OG
Rec	SPEN	42	4.15e-13	0	51.4	TSG
Rec	CDKN1B	11	7.5e-13		100	TSG
Dom	SF3B1	19	1.48e-11	60	0	OG
	ARHGAP35	16	6.62e-7	0	64.3	TSG

Showing 11 to 20 of 104 entries

Previous 1 2 3 4 5 … 11 Next

Fig. 4. SomInaClust prediction of candidate driver genes. A. Pyramid plot showing Oncotype DX genes of which were predicted as candidate driver genes with calculated oncogene (OG) and tumor-suppressor (TSG) scores by SomInaClust R package. B. Detailed results of SomInaClust analysis with number of mutations, OG score, TSG score and q-value (qDG). CGC: COSMIC cancer gene census, Rec: Recessive (TSG), Dom: Dominant (OG).

Transcriptome analysis module provides differential expression analysis (DEA) of RNAseq data by comparing the expression of genes in primary tumor samples against adjacent normal samples. We present two results options for this analysis: ‘Paired’ as comparison of tumor samples against their own paired normal or ‘All’ as comparison of tumor samples against a normal sample subset of patients if such is available for the particular cancer. Volcano plot in Figure 5A summarizes the differential expression analysis of paired BRCA samples and Oncotype DX genes highlighted through the ‘My Genes’ panel. Table in Figure 5B presents the details of DEA with gene symbol, fold change (logFC) and p values of top 10 significantly differentially expressed genes ranked by p-value.



SYMBOL	ensembl_gene_id	logFC	AveExpr	t	P.Value	adj.P.Val	B
COL10A1	ENSG00000123500	7.5182191481065	3.28051492355376	30.1100786176065	4.91127617127651e-81	9.13595593380857e-77	174.198697361669
KLHL29	ENSG00000119771	-2.92903062019832	3.24074313563804	-29.5220001328041	1.72154114140026e-79	1.60120541561639e-75	170.750267525278
SPRY2	ENSG00000136158	-2.63285717689352	4.72370365156901	-29.2625524349796	8.37947024083717e-79	5.19583018066843e-75	169.25888850825
VEGFD	ENSG00000165197	-5.66324833308504	1.99578093853064	-28.9317591230961	6.37822650954394e-78	2.96619423826341e-74	167.058798563458
MMP11	ENSG00000099953	6.07902876417296	4.79341497955916	28.883225937226	8.60041471246827e-78	3.1996982896267e-74	166.878495090989
ADAMTS5	ENSG00000154736	-3.11644577620304	5.23042731699242	-27.8706098659734	4.69822162937804e-75	1.45660531249484e-71	160.657219383908
PAMR1	ENSG00000149090	-3.83381622141784	4.08287279436218	-26.4614519291141	3.74685910366137e-71	9.95701043518697e-68	151.682874961013
CAPN11	ENSG00000137225	-3.02645813787852	1.25950345569362	-26.3637431378098	7.05049347818897e-71	1.63941599601589e-67	150.909124125522
TMEM220	ENSG00000187824	-2.33758621319281	2.75154492057195	-26.2506847477051	1.46741730923573e-70	3.03298853182256e-67	150.2846042557
PPP1R14A	ENSG00000167641	-3.40601652538607	2.34679469949561	-26.0628288591903	4.97779777740038e-70	9.25969942552019e-67	149.045581545073

Showing 1 to 10 of 18,602 entries

Previous 1 2 3 4 5 … 1,861 Next

bioRxiv preprint doi: https://doi.org/10.1101/2023.01.20.524925; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig. 5. Differential expression of genes in tumor samples versus normal samples. A. Volcano plot showing up-regulated and down-regulated genes with -log10 conversion of p-values. Oncotype DX genes are highlighted on the graph. B. Differential expression results table presenting gene symbols, fold changes (logFC) and adjusted-p-values. C. Heatmap showing pathway enrichment of differentially expressed genes.

Pathway enrichment of differentially expressed genes showed that these genes play role in focal adhesion and ECM-receptor interaction which can be related with metastasis; Ras signaling, PI3K-Akt signaling, cAMP signaling and Phenylalanine metabolism pathways which are related with cell growth (Figure 5C). Genes related to these pathways can be observed with their p value color representation at heat plot of pathway enrichment analysis (Figure 5C).

Metastasis related gene MMP11, proliferation related genes BIRC5, MYBL2, MKI67 (Ki67), AURKA (STK15), CCNB1 and ERBB2 (HER2) from Oncotype DX gene set are highly up-regulated significantly in tumor samples of BRCA (Figure 5A). However hormone related genes (BAG1, BCL2, CD68, ESR1 (ER), GSTM1, PGR, SCUBE2) are not significantly differentially expressed among all tumor samples.

When we concentrate on the ERBB2 gene because it was predicted as a driver oncogene, we can visualize positions of mutations by the Lollipop plot in Figure 6A. Most of the mutations of ERBB2 gene are located on the kinase domain of HER2 (Figure 6A). These mutations are mostly missense on protein positions 755 (n=7), 767 (n=2), 769 (n=3), 777 (n=4), 797 (n=1), 842 (n=1), 939 (n=1) and in frame insertion on protein position 885 (n=1). Mutations on ERBB2 gene in tumor samples cause lower survival probability with 1.43 hazard ratio (not significant, p=0.084) (Figure 6B).

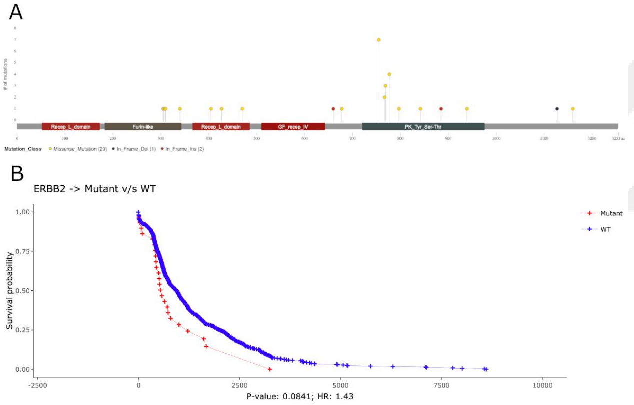


Fig. 6. Detailed analysis of ERBB2 (HER2) mutations. A. Lollipop plot showing mutations of ERBB2 gene among tumor samples. B. Overall survival analysis of wild type versus mutated ERBB2 in tumor samples.

When we checked the expression levels of ERBB2 in tumor samples versus normal samples, from paired DEA, ERBB2 is expressed in tumor samples significantly higher than their adjacent normal samples (p=3.521E-10) (Figure 7A). However, patients with higher expression of ERBB2 have higher survival probability significantly (p=0.045) (Figure 7B).

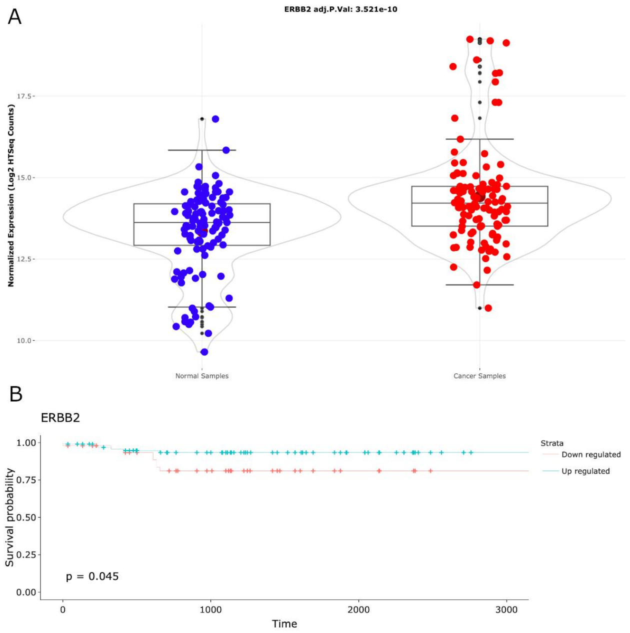
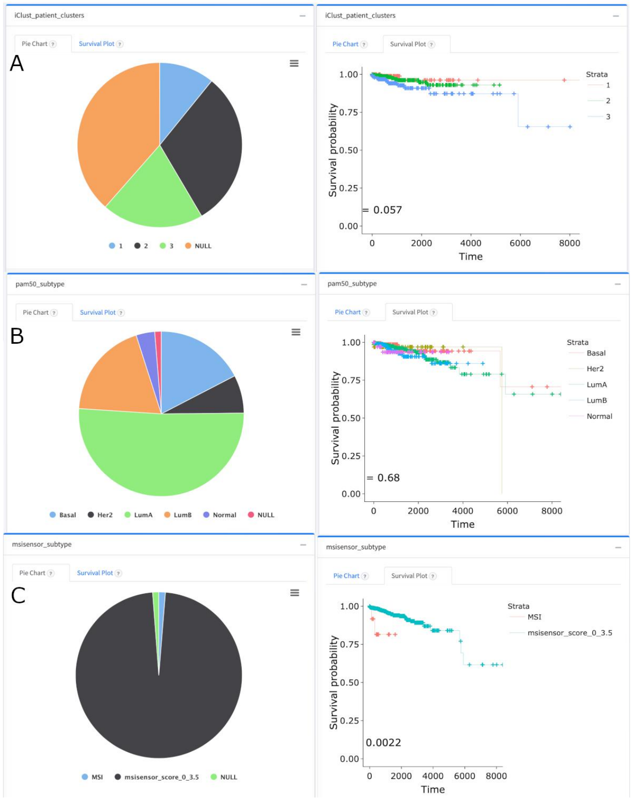


Fig. 7. Detailed analysis of ERBB2 (HER2) expression. A. Violin plot presenting log2 transformed normalized mRNA expression of ERBB2 in normal and tumor samples with adjusted p-value. B. Overall survival analysis of expression levels of ERBB2 in tumor samples.

Clinical data analysis is composed of pie chart visualization and survival analysis of clinical features with curated patient clusters and sample subtypes. Figure 8 shows the visualization of iClusters, PAM50 subtypes and MSI-sensor subtypes. iClusters showed differential survival probability close to significance level (p=0.057) (Figure 8A), however PAM50 subtypes do not have differential survival probabilities. (p=0.68) (Figure 8B). Patients who have tumors with MSI have lower survival probability than patients with MSI stable tumors (p=0.0022) (Figure 8C).



bioRxiv preprint doi: https://doi.org/10.1101/2023.01.20.524925; this version posted January 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig. 8. Pie charts and survival analysis of clinical features. A. Pie chart representation and survival analysis of iClusters. B. Pie chart representation and survival analysis of PAM50 subtypes. C. Pie chart representation and survival analysis of MSI subtypes.

Radial slices of the pie-chart are clickable, letting the user add the corresponding patient subsets to the 'My Patients' clipboard panel. Besides, users can customize a variety of plots such as survival plot, volcano plot, box plot, heatmaps, lollipop plot, and pie charts for discovering common molecular profiles for precision oncology. Each plot and data table are downloadable for use in articles.

4 Conclusion

Several web portals facilitating analysis on TCGA data have been developed and widely used such as Genomic Data Commons (GDC) data portal (Grossman, 2016), ICGC data portal (Zhang, 2019) and CPTAC data portal (Edwards, 2015). The cBioPortal is an open-access, open-source resource for interactive exploration of multidimensional cancer genomics data sets (Cerami, 2012; Gao, 2013) providing gene-centered query and visualization functions across multiple cancers. IntOGen is another similar framework for automated comprehensive knowledge extraction based on mutational data from sequenced tumor samples from TCGA patients (Francisco, 2020). However, we provide pre-performed SNV, CNV and differentially expression analyses with large sets of patient clusters and sample subtype information. We present signature based clustering using Generalized Linear Model for three cancer types (LUAD, LUSC and COAD). For all 33 cancer types immune and MSI-sensor scores of all patients are retrieved from their original publications. For the breast cancer (BRCA), PAM50 and TNBC patient cohorts are retrieved from their original publications. For fifteen cancer types, previously published TCGA cohorts of the individual tumor types are retrieved by curatedTCGAData R package (Ramos et al, 2020). By the time this manuscript is written iClusterPlus based patient cohorts are generated for fifteen cancers based on five data dimensions: miRNA, methylation, single nucleotide variation, transcriptome and copy number variation. A re-runnable parallel Linux pipeline is implemented enabling a scalable update of the pan-cancer data at the backend. We plan to generate iClusters for 33 cancer types, and integrate results of miRNA and methylation analyses, too. Since its initial inception to public in January 1st of 2022, TCGAnalyzeR has been regularly accessed by ~79 unique users/day.

TCGAnalyzeR provides a user-friendly web framework for integrative, large-scale analyses of genomic and clinical data of 33 cancer types from TCGA. TCGAnalyzeR web interface allows clinical oncologists and cancer researchers to create subcohorts and/or gene sets of interest to filter visualization of analyses. TCGAnalyzeR help page includes a demonstration of the app with the two use-cases of subcohort discovery and can be used as a manual.

Funding

This work has been supported by Turkish National Institutes of Health (TÜSEB) [4583].

Conflict of Interest: none declared.

References

Berger AC, Korkut A, Kanchi RS, Hegde AM, Lenoir W, Liu W, Liu Y, Fan H, Shen H, Ravikumar V, Rao A, Schultz A, Li X, Sumazin P, Williams C, Mestdagh P, Gunaratne PH, Yau C, Bowlby R, Robertson AG, Tiezzi DG, Wang C, Cherniack AD, Godwin AK, Kuderer NM, Rader JS, Zuna RE, Sood AK, Lazar AJ, Ojesina AI, Adebamowo C, Adebamowo SN, Baggerly KA, Chen TW, Chiu HS, Lefever S, Liu L, MacKenzie K, Orsulic S, Roszik J, Shelley CS, Song Q, Vellano CP, Wentzensen N; Cancer Genome Atlas Research Network; Weinstein JN, Mills GB, Levine DA, Akbani R. A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. Cancer Cell. 2018 Apr 9;33(4):690-705.e9. doi: 10.1016/j.ccell.2018.03.014. Epub 2018 Apr 2. PMID: 29622464; PMCID: PMC5959730.

Carey V (2022). _BiocOncoTK: Bioconductor components for general cancer genomics_. R package version 1.18.0.

Cerami, E. et al. (2012) The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. Cancer Discov., 2, 401–404.

Chakravarty, D. et al (2017). OncoKB: A Precision Oncology Knowledge Base. JCO Precis Oncol. 2017 Jul;2017:PO.17.00011. doi: 10.1200/PO.17.00011. Epub 2017 May 16.

Colaprico,A. et al. (2016) TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res., 44, e71–e71.

Deng,M. et al. (2017) FirebrowseR: an R client to the Broad Institute's Firehose Pipeline. Database, 2017, baw160.

Ding L, Bailey MH, Porta-Pardo E, Thorsson V, Colaprico A, Bertrand D, Gibbs DL, Weerasinghe A, Huang KL, Tokheim C, Cortés-Ciriano I, Jayasinghe R, Chen F, Yu L, Sun S, Olsen C, Kim J, Taylor AM, Cherniack AD, Akbani R, Suphavilai C, Nagarajan N, Stuart JM, Mills GB, Wyczalkowski MA, Vincent BG, Hutter CM, Zenklusen JC, Hoadley KA, Wendl MC, Shmulevich L, Lazar AJ, Wheeler DA, Getz G; Cancer Genome Atlas Research Network. Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. Cell. 2018 Apr 5;173(2):305-320.e10

Durinck,S. et al. (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc., 4, 1184–1191.

Galili,T. et al. (2018) heatmaply: an R package for creating interactive cluster heatmaps for online publishing. Bioinformatics, 34, 1600–1602.

Gao,J. et al. (2013) Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. Sci. Signal., 6, pl1–pl1.

Goldman,M.J. et al. (2020) Visualizing and interpreting cancer genomics data via the Xena platform. Nat. Biotechnol., 38, 675–678.

Grossman,R.L. et al. (2016) Toward a Shared Vision for Cancer Genomic Data. N. Engl. J. Med., 375, 1109–1112.

Guo,X. et al. (2019) G3viz: an R package to interactively visualize genetic mutation data using a lollipop-diagram. Bioinformatics, btz631.

Kassambara A. (2022) survminer: Drawing Survival Curves using 'ggplot2'

Kunst,J. (2022) highcharter: A Wrapper for the 'Highcharts' Library.

Lawrence,M. et al. (2013) Software for Computing and Annotating Genomic Ranges. PLOS Comput. Biol., 9, e1003118.

Lehmann BD, Jovanović B, Chen X, Estrada MV, Johnson KN, Shyr Y, Moses HL, Sanders ME, Pietenpol JA. Refinement of Triple-Negative Breast Cancer Molecular Subtypes: Implications for Neoadjuvant Chemotherapy Selection. PLoS One. 2016 Jun 16;11(6):e0157368. doi: 10.1371/journal.pone.0157368. PMID: 27310713; PMCID: PMC4911051.

Mo,Q. et al. (2018) A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. Biostatistics, 19, 71–86.

Mo,Q. and Shen,R. (2022) iClusterPlus: Integrative clustering of multi-type genomic data.

Morganella,S. et al. (2011) Finding recurrent copy number alterations preserving within-sample homogeneity. Bioinformatics, 27, 2949–2956.

Ramos,M. et al. (2020) Multiomic Integration of Public Oncology Databases in Bioconductor. JCO Clin. Cancer Inform., 958–971.

Ritchie,M.E. et al. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res., 43, e47–e47.

Robinson,M.D. et al. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26, 139–140.

Sakaguchi T, Iketani A, Furuhashi K, et al. Comparison of the analytical performance between the Oncomine Dx Target Test and a conventional single gene test for epidermal growth factor receptor mutation in non-small cell lung cancer. Thorac Cancer. 2021;12(4):462-467. doi:10.1111/1759-7714.13767

The Cancer Genome Atlas Research Network. et al. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet., 45, 1113–1120.

Therneau,T. (2022) A Package for Survival Analysis in R.

Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, Porta-Pardo E, Gao GF, Plaisier CL, Eddy JA, Ziv E, Culhane AC, Paull EO, Sivakumar IKA, Gentles AJ, Malhotra R, Farshidfar F, Colaprico A, Parker JS, Mose LE, Vo NS, Liu J, Liu Y, Rader J, Dhankani V, Reynolds SM, Bowlby R, Califano A, Cherniack AD, Anastassiou D, Bedognetti D, Mokrab Y, Newman AM, Rao A, Chen K, Krasnitz A, Hu H, Malta TM, Noushmehr H, Pedamallu CS, Bullman S, Ojesina AI, Lamb A, Zhou W, Shen H, Choueiri TK, Weinstein JN, Guinney J, Saltz J, Holt RA, Rabkin CS; Cancer Genome Atlas Research Network; Lazar AJ, Serody JS, Demicco EG, Disis ML, Vincent BG, Shmulevich I. The Immune Landscape of Cancer. Immunity. 2018 Apr 17;48(4):812-830.e14. doi: 10.1016/j.immuni.2018.03.023. Epub 2018 Apr 5. Erratum in: Immunity. 2019 Aug 20;51(2):411-412. PMID: 29628290; PMCID: PMC5982584.

Van den Eynden,J. et al. (2015) SomInaClust: detection of cancer genes based on somatic mutation patterns of inactivation and clustering. BMC Bioinformatics, 16, 125.

Wickham,H. et al. (2022) readr: Read Rectangular Text Data.

Yu,G. et al. (2012) clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. OMICS J. Integr. Biol., 16, 284–287.

Zengin,T. and Önal-Süzek,T. (2020) Analysis of genomic and transcriptomic variations as prognostic signature for lung adenocarcinoma. BMC Bioinformatics, 21, 368.

Zengin,T. and Önal-Süzek,T. (2021) Comprehensive Profiling of Genomic and Transcriptomic Differences between Risk Groups of Lung Adenocarcinoma and Lung Squamous Cell Carcinoma. J. Pers. Med., 11, 154.

CURRICULUM VITAE

T** Z*******
D********* of M******* B****** and G*******
F****** of S******
M**** S**** K***** U*********

Education

Degree	Institution	Graduation year
Bachelor's Degree	Izmir Institute of Technology Department of Molecular Biology and Genetics	2011
Master's Degree	Izmir Institute of Technology Department of Molecular Biology and Genetics	2014
Master's Degree	Muğla Sıtkı Koçman University Department of Bioinformatics	2019
Doctorate	Muğla Sıtkı Koçman University Department of Bioinformatics	2023

Projects

National Projects

Target Specific Pan-Cancer Therapies (PAN-TER). TÜBİTAK 1004, 2021-2024.

Development of a web platform for cancer drug repurposing. TÜSEB - Systems Biology and Bioinformatics Strategic R&D Project Call (2019-TA-01), 2020-2023.

Investigation of the Determinant Role of IRF6 in the Inhibitory or Stimulatory Effects of Notch Pathway on Cell Division. TÜBİTAK, 2011-2014.

International Projects

Peritoneal Immune Modulation for Colorectal Peritoneal Metastases. EU, TRANSCAN, 2023-2026.

Research Abroad

Antibody-antigen molecular docking. University of Oslo, Norway, 2021.

Curation of antibody-antigen crystal complexes. University of Oslo, Norway, 2022.

Publications

Talip Zengin, Başak Abak Masud, Tuğba Önal-Süzek. TCGAnalyzeR: a web application for integrative visualization of molecular and clinical data of cancer patients for cohort and associated gene discovery. bioRxiv, 2023.

Rahmad Akbar, Habib Bashour, Puneet Rawat, Philippe A. Robert, Eva Smorodina, Tudor Stefan Cotet, Karine Flem-Karlsen, Robert Frank, Brij Bhushan Mehta, Mai Ha Vu, **Talip Zengin**, Jose Gutierrez-Marcos, Fridtjof Lund-Johansen, Jan Terje Andersen & Victor Greiff. Progress and challenges for the machine learning-based design of fit-for-purpose monoclonal antibodies. mAbs, 2022.

Mehmet Cihan Sakman, **Talip Zengin**, Deniz Kurşun, Tuğba Süzek. A Personalized Oncology Mobile Application Integrating Clinical and Genomic Features to Predict the Risk Stratification of Lung Cancer Patients via Machine Learning. Mugla Journal of Science and Technology, 2022.

Talip Zengin, Tuğba Önal-Süzek. Comprehensive Profiling of Genomic and Transcriptomic Differences between Risk Groups of Lung Adenocarcinoma and Lung Squamous Cell Carcinoma. Journal of Personalized Medicine, 2021.

Talip Zengin, Tuğba Önal-Süzek. Analysis of genomic and transcriptomic variations as prognostic signature for lung adenocarcinoma. BMC Bioinformatics, 2020.

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

International Symposiums

Talip Zengin, Tuğba Süzek, 2019. A Meta-Analysis of Genomic and Transcriptomic Variations in Lung Adenocarcinoma. The 6th International Workshop on Computational Network Biology: Modeling, Analysis, and Control (CNB-MAC 2019)

Talip Zengin, Tuğba Süzek, 2019. Meta-Analysis Pipeline for Genomics and Transcriptomics Variations in TCGA Data. The 12th International Symposium on Health Informatics and Bioinformatics.

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

Süleyman Arzıman, **Talip Zengin**, Ömür Baysal, Kubilay Kurtuluş Baştaş, 2017. In-silico analysis on gene sequence encoding type III effector proteins reflect similar host- speciation for bacterial pathogens infecting tomato. 4th ICSAE

Talip Zengin, Özden Yalçın Özuysal, 2013. IRF6 is a downstream mediator of Notch signaling in breast epithelial cells. VII. Notch Meeting

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

Book Chapter

Emre Balta, Seçkin Boz, Umut Rende, **Talip Zengin**, Işıl Erbaşol 2012. BLAST. Biyoenformatik I - Dizi Kıyaslamaları. Editor: Jens Allmer, Canan Has, Şule Yılmaz. Nobel Akademik Yayıncılık. Ankara.