

Identification and Functional Analysis of Hub Genes in Cell Cycle Pathway Linked to the Development of Lung Cancer with Non-Small Cells

Meixi Chen^{1,*}, Jiajing Pan², Guanyang Wang³

¹School of Life Sciences, Shanghai University, Shanghai, China

²School of Basic Medicine, Fujian Medical University, Fujian, China

³School of Horticulture, Nanjing Agricultural University, Nanjing, China

*Corresponding author: chenmeixi@shu.edu.cn

Abstract. Non-small cell lung cancer (NSCLC) is the most prevalent cause of cancer-related fatalities. We combined different gene expression datasets to screen out eight hub genes (CDC6, PLK1, BUB1B, CHEK1, TTK, CDC20, CCNB2, and PTTG1) associated with the Cell cycle pathway, which also showed significant poor prognosis and late upregulation of LUAD in NSCLC. KRAS and EGFR are the star molecules in lung cancer. KRAS causes overexpression of hub genes through up-regulation of the cell cycle pathway, and the up-regulation of hub genes in NSCLC can be confirmed at amounts of mRNA and proteins. Hub genes have a certain correlation with EGFR, and upregulation of the cell cycle pathway may cause resistance in NSCLC caused by EGFR mutations. The emergence of this resistance may be related to the change of Gene copy number alteration (CAN) and co-occurrence of BUB1B and CCNB2, PLK1 and TTK, CHEK1 and TTK. These cell cycle subtype genes provide a new perspective for treating patients with NSCLC, and the specific mechanism of hub genes can be further studied.

Keywords: non-small cell lung cancer; Cell cycle; bioinformatics analysis; KRAS mutation; EGFR mutation; Gene Copy Number Alteration.

1. Introduction

One of the main causes of cancer-related fatalities globally is lung cancer [1, 2]. Out of all the numerous kinds of lung cancer, non-small cell lung cancer (NSCLC) is the most typical variety and has a worst prognosis [3]. Understanding the molecular mechanisms underlying NSCLC development and progression is crucial for the creation of successful treatment plans.

Dysregulation of the cell cycle is a major factor in the onset and spread of cancer. The cell cycle is a highly regulated process involving various checkpoints and signaling pathways that ensure the accurate replication and division of cells [4]. Alterations in key genes involved in the cell cycle pathway have been connected to the aetiology of multiple cancers, comprising NSCLC.

One such important pathway involved in NSCLC is the KRAS and EGFR signaling pathways. KRAS and EGFR are well-known oncogenes that play crucial roles in cell proliferation, survival, and tumor progression. Mutations or dysregulation of these genes have been connected to the emergence and development of NSCLC. Furthermore, studies have shown that the dysregulation of the cell cycle pathway, mediated by KRAS and EGFR, contributes to the resistance observed in NSCLC patients [5].

However, there are still some limitations while providing therapeutic care for NSCLC. These include the development of treatment resistance, challenges in selecting the optimal treatment strategy, individual variations in treatment response, and considerations regarding treatment side effects and safety. Further research is required to understand resistance mechanisms, develop personalized treatment strategies, and enhance side effect management and safety monitoring. Although limitations exist, we can anticipate more innovations and progress in the field with the continuous advancement of science and technology, offering more effective treatment options for NSCLC patients [6].

The purpose of this work is to look at how cell cycle dysregulation functions in NSCLC and how it relates to EGFR and KRAS. We have selected eight hub genes associated with the cell cycle pathway, showing significant upregulation and adverse prognosis in NSCLC patients. By integrating different gene expression datasets, we have identified these hub genes and confirmed their dysregulation at both mRNA and protein levels. Additionally, we will explore the potential relationship between these hub genes and the KRAS and EGFR signaling pathways.

Understanding the specific mechanisms underlying cell cycle dysregulation and its interaction with the KRAS and EGFR pathways in NSCLC holds great promise for the development of targeted therapies and the improvement of patient outcomes. By elucidating the complex interplay between these molecular pathways, we can potentially identify novel therapeutic targets and strategies to overcome drug resistance in NSCLC.

Overall, this study aims to provide new insights into the molecular mechanisms driving NSCLC development and progression, with a particular focus on cell cycle dysregulation and its association with the KRAS and EGFR pathways (Fig. 1). The findings from this study have the potential to contribute significantly to the development of personalized methods of therapy for NSCLC patients, eventually becoming better their prognosis as well as living quality.

2. Materials and Methods

2.1. Data from Gene Expression Profiles

The Gene Expression Omnibus (GEO) database provided the microarray dataset GSE19188. The dataset included a total of 156 tumour and normal samples from individuals diagnosed with early-stage NSCLC between November 2009 and March 2019. The RNA sequencing data obtained from the TCGA (The Cancer Genome Atlas) were acquired in their original, unprocessed format. A cohort of 1017 lung cancer samples and 110 corresponding examples of nontumors were procured for more examination.

2.2. Identification of DEGs

Aiming to determine differentially expressed genes (DEGs) within the validation cohort of the GEO NSCLC dataset, the LIMMA package (LIMMA) in R software was employed. For filtering the DEGs, The cut offs were set at false discovery rate (p-value) < 0.05 and $|\log_2 \text{fold change (FC)}| \geq 1$. We used the edgeR Bioconductor project programme (edgeR) to analyse the validation cohort from the TCGA LUAD and LUSC dataset. P-values < 0.01 and $|\log_2 \text{FC}| \geq 2$ were regarded as substantially differentially expressed genes. For future study, the DEGs that overlapped TCGA and GEO datasets were retained.

2.3. Protein Protein Interaction (PPI) Web and Upregulation in the Kyoto Encyclopaedia of Genes and Genomes (KEGG)

The Database of Annotation, Visualisation, and Integrated Discovery (DAVID) was utilised for KEGG dissection. One widely used online database is the PPI web, which was constructed using the Search Tool for the Retrieval of Interacting Genes (STRING).

2.4. DEGs' overall survival (OS)

The article used the Kaplan Meier plotter (<https://kmplot.com/analysis/>) for survival analysis. GEO and TCGA provided the gene expression and OS data (Gyorffy et al., 2013). The tumour and healthy tissues were split into two categories according to median expression so as to assess the prognostic key genes' significance. The Kaplan Meier plotter accustomed to create OS curves, moreover, the outcomes included the 95% confidence intervals (CIs) for the hazard ratio (HR) and a p value for log rank. An online server called Gene Expression Profiling Interactive Analysis (GEPIA) was used to carry out the Kaplan-Meier approach.

2.5. Pathology Analysis of Core Genes

The Human Protein Atlas (HPA) database (<http://www.proteinatlas.org/>) provided the immunohistochemistry (IHC) staining pictures of CDC6, PLK1, TTK, CDC20, CCNB2, and PTTG1 from lung cancer and normal tissues for further research. Image analysis is performed using ImageJ software.

2.6. Gene Set Enrichment Analysis

To find out whether there were any notable variations in the discovered genes' expression between the two groups, Gene Set Enrichment Analysis was used. Target sets, Hallmark gene sets v7.2, were downloaded, and GSEA (Subramanian et al., 2005) was carried out with the use of GSEA_4.0.3 software. The statistical measures of nominal p value, false discovery rate (FDR), and normalised enrichment score (NES) were used to detect the pathways that exhibited enrichment in each phenotype. Gene sets were deemed significant if they had a NOM $p < 0.05$.

2.7. Statistical Analyses

GraphPad Prism 9 and R software (version 4.3.1), were applied to every statistical study. For every statistical test, $p < 0.05$ was the cutoff point for statistical significance.

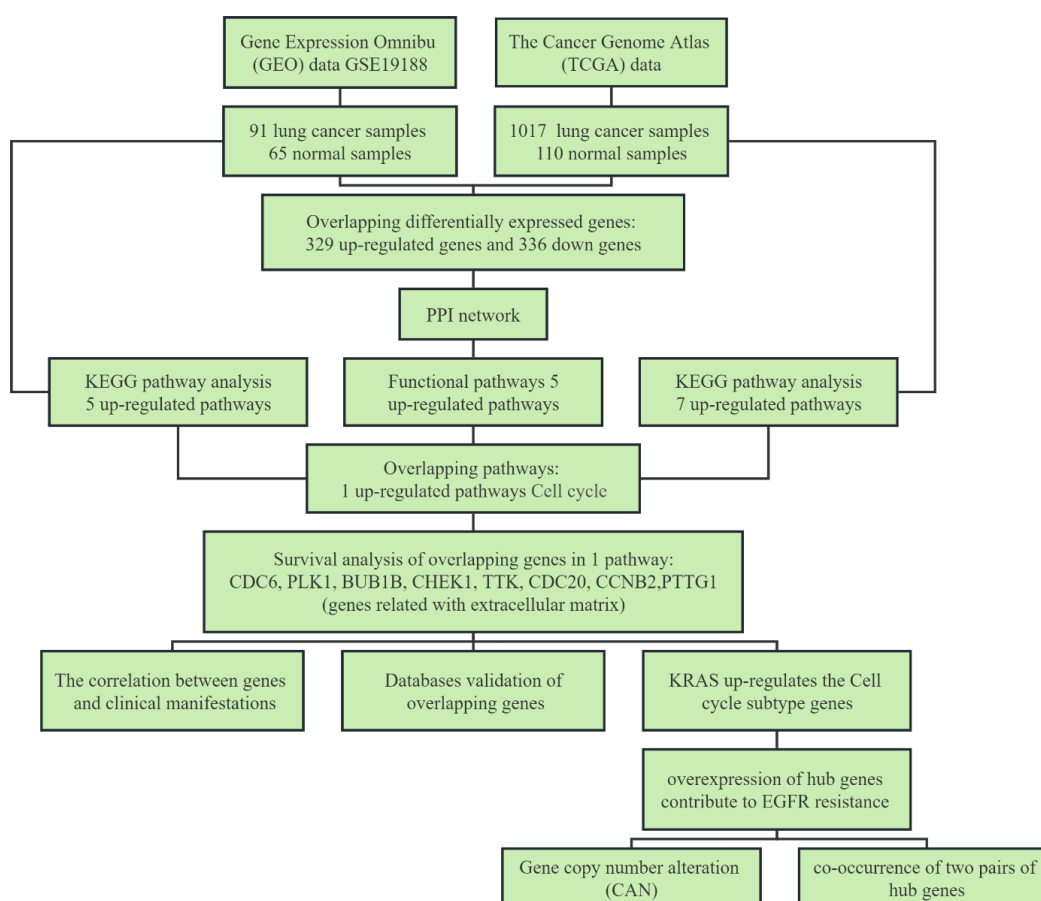


Fig. 1 Diagram illustrating the flow of the methodical, thorough study of DEGs from patients in the GEO series and TCGA groups, as well as the relationship between DEG expression and overall survival (OS). Additional research showed that the cell cycle signature controls the development of lung cancer

3. Results

3.1. Tumour Gene Expression Programmes Represent Clinically Distinct Cell Cycle Subtypes

3.1.1 DEGs Discovered Using NSCLC TCGA and GEO Profiles

Our goals were to determine clinically meaningful molecular subgroups, find important markers for NSCLC, and find the DEGs between healthy and tumour tissues. First, DEGs were identified using RNA-Seq data from the TCGA and GEO databases. The PPI network and KEGG pathway studies were performed using the overlapped DEGs from two datasets. One common up-regulated route was filtered by further analysing the highest 7 pathway activation from the GEO and TCGA datasets, the KEGG pathway, and the PPI web. Hub genes were identified as eight genes (CDC6, PLK1, BUB1B, CHEK1, TTK, CDC20, CCNB2, and PTTG1) that were shared by the pathway.

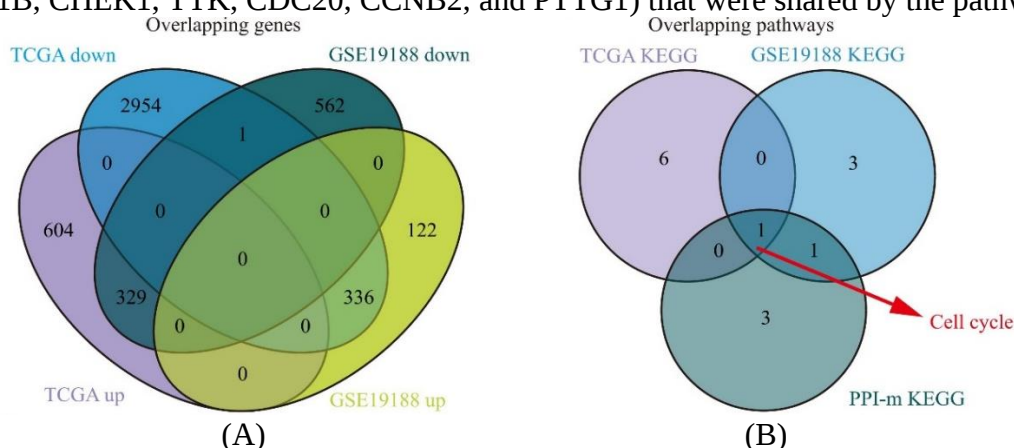


Fig. 2 (A) This study use Venn diagrams to illustrate the overlap of differentially expressed genes (DEGs) across two datasets: GSE19188 and The Cancer Genome Atlas (TCGA) (B) Venn charts illustrating the intersecting routes

3.1.2 PPI Web and Module Analysis Used to Identify Hub Genes

The TCGA LCAD and LUSC datasets, together with the GSE19188 gene expression profile, were used to determine the common critical genes and pathways that could be crucial in the occurrence of lung cancer. The GSE19188 dataset, which includes 65 matched neighbouring normal samples and 91 lung tumour samples, yielded a total of 869 DEGs. There were 458 upregulated and 892 downregulated genes among these 1350 DEGs. There were 3291 upregulated and 933 downregulated genes in the DEGs that were obtained from the TCGA LCAD and LUSC dataset. The GSE19188 and TCGA LCAD and LUSC datasets included a grand total of 665 DEGs, encompassing 336 higher-regulated genes and 329 downregulated genetic variants (Fig. 2A). Using the DEGs from each dataset, A KEGG pathway analysis was done to look into the possible biological implications of these genes. Seven pathways were shown to enrich the group of people with lung cancer, according to KEGG analysis (Fig. 2B). The DEGs of the GSE19188 collection were shown to be highly enriched in 5 pathways, involving the cell cycle, motor proteins, p53 signaling pathway, Cellular senescence, and Oocyte meiosis (Fig. 3A). The top 7 enhanced pathways that are elevated from TCGA LCAD and LUSC statistic comprised Neuroactive ligand–receptor interaction, Alcoholism, Cell cycle, Neutrophil extracellular trap formation, Systemic lupus erythematosus, Protein digestion and absorption and Biosynthesis of amino acids (Fig. 3B). The PPI web of DEGs constructed with Cytoscape and the String software is depicted in Fig. 4A, and Fig. 4B identifies the main module of the PPI web based on the score of each node.

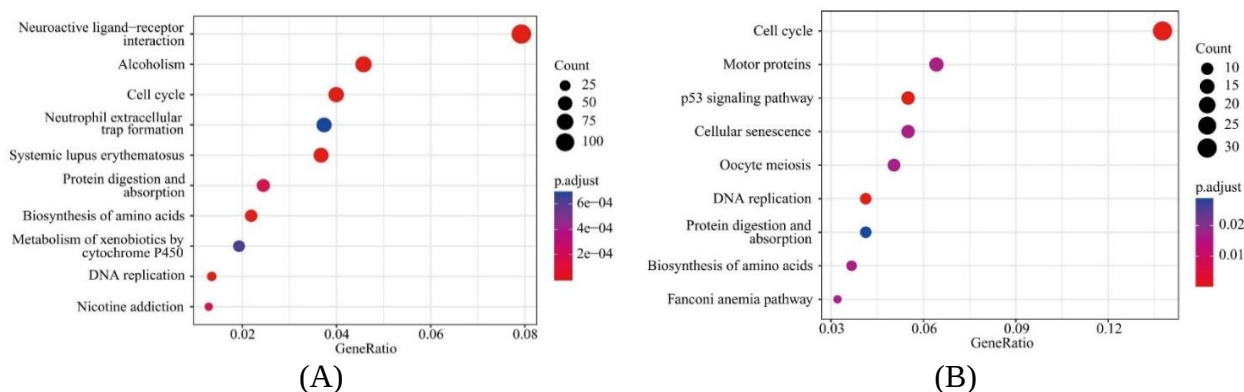


Fig. 3 (A) The DEGs in the GSE19188 dataset were subjected to a pathway enrichment analysis using the KEGG (B) The TCGA dataset's DEGs experienced the investigation of KEGG pathway enrichment

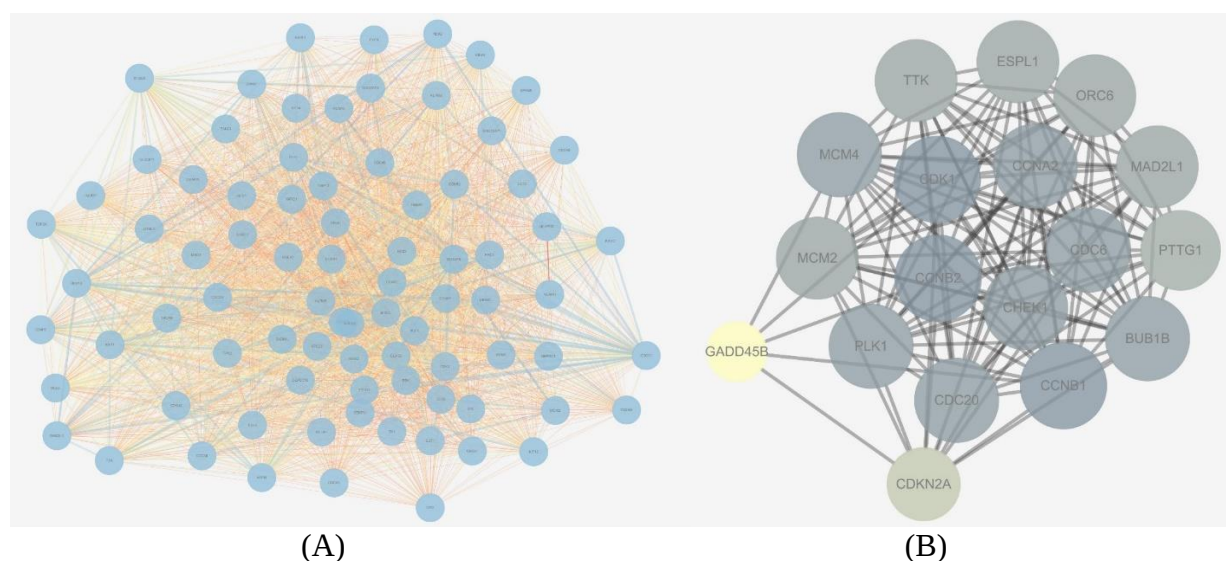


Fig. 4 (A) PPI web of DEGs created using Cytoscape and the String programme (B) The PPI web's primary module, determined by each node's score

3.2. A Negative Prognosis and Advanced Clinicopathological Indicators Correlated with Elevated Cell Cycle Subtype Gene Expression

3.2.1 Analysis of Hub Gene Survival in NSCLC

Kaplan-Meier Plotter (<http://kmplot.com/>) enables survival analysis of pivot genes with worse OS (Fig. 5) to further check the link with the NSCLC clinical patients. The p-values of all core genes are less than $1e-16$, demonstrating some confidence in the survival assays obtained. The best results in survival analysis were CDC6 [HR = 2.43 (2.06–2.86), $p < 1e-16$], CCNB2 [HR = 2.36 (2.01–2.77), $p < 1e-16$] and PTTG1 [HR = 2.11 (1.8–2.46), $p < 1e-16$]. However, PLK1 [HR = 1.74 (1.54–1.96), $p < 1e-16$] and BUB1B [HR = 1.96 (1.69–2.27), $p < 1e-16$] were less obvious.

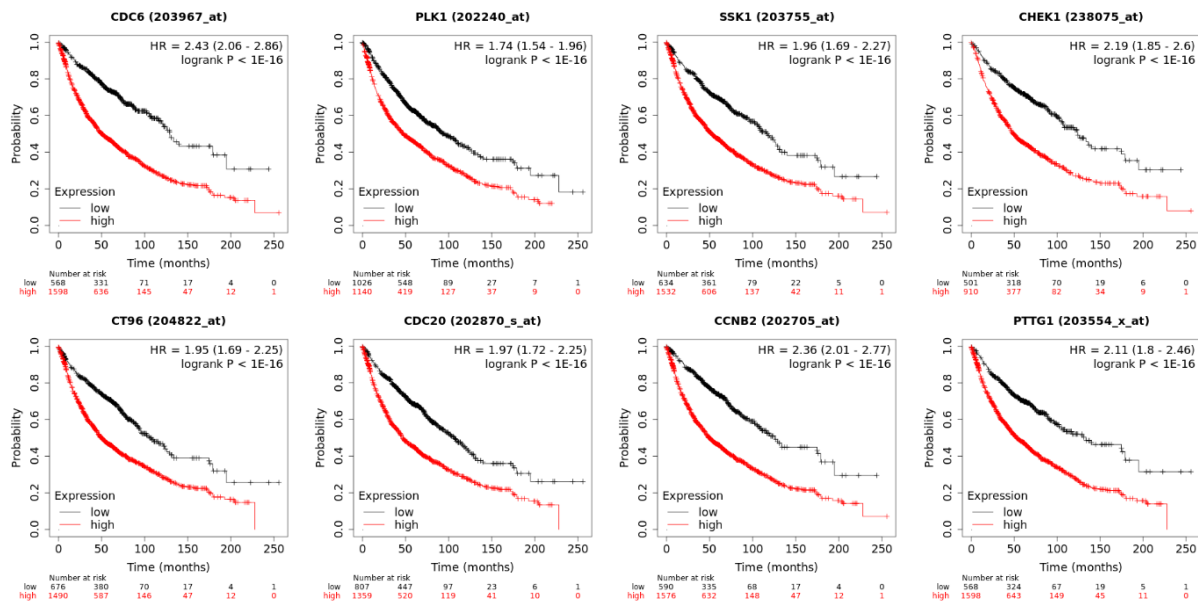
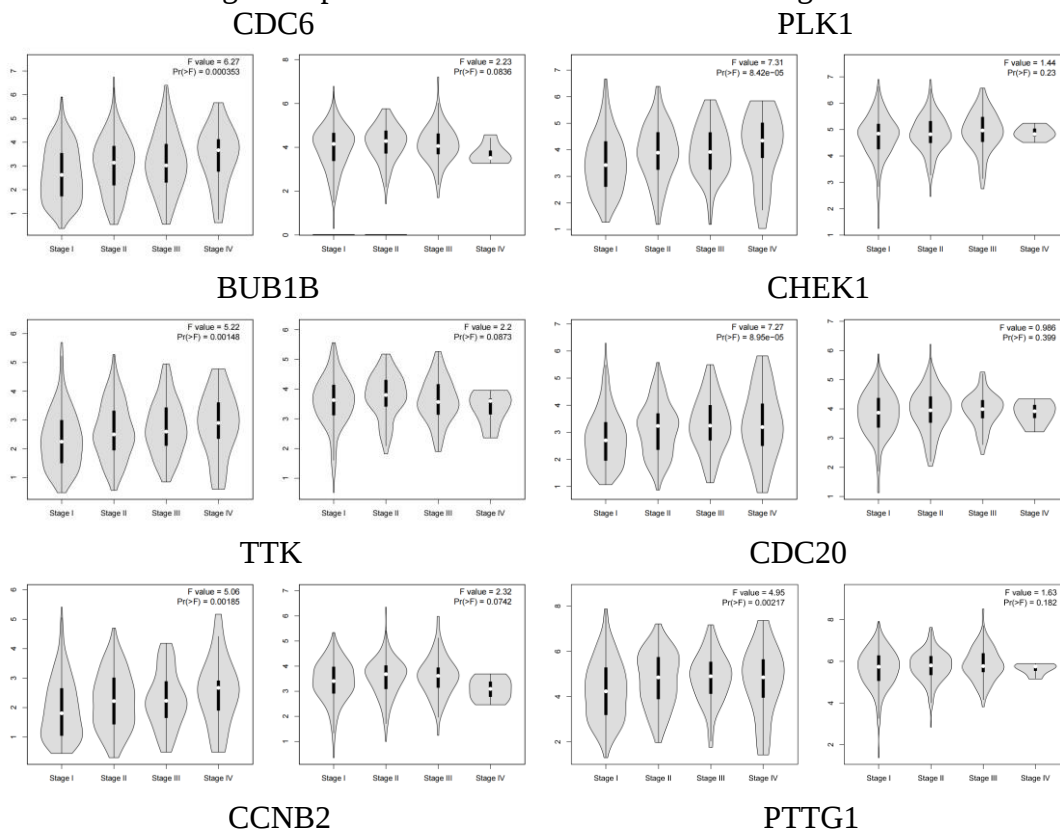


Fig. 5 Survival analysis for hub genes CDC6, PLK1, TTK, CDC20, CCNB2 and PTTG1

3.2.2 Pathological Stage Plot of hub genes in NSCLC

Subsequently, an inquiry was made on the extent to which each of the aforementioned seven genes had a strong association with the clinical traits of the individuals. As stated by the data shown in Fig. 6, there was a notable increase in the expression levels of CDC6, PLK1, BUB1B, CHEK1, TTK, CDC20, CCNB2, and PTTG1 in patients with advanced clinical stage of lung adenocarcinoma (LUAD) as compared to those with early stage LUAD. On the contrary, there is no apparent association between hub gene expression in LUSC and the clinical stage.



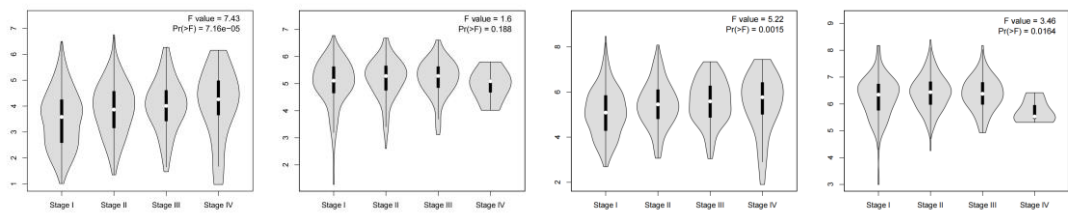


Fig. 6 the connection between the clinical stage and hub genes. There was LUAD on the left and LUSC on the right

3.3. Research on Hub Gene Upstream Regulatory Mechanisms

The primary mutations causing lung cancer to arise are KRAS mutations. Greater amounts of hub genes (CDC6, BUB1B, CHEK1, TTK, CDC20, CCNB2, and PTTG1) were seen Among people with KRAS mutations who have lung cancer, according to our analysis of cohorts of lung disease patients Having mutations in KRAS (GSE31210) (Fig. 7). Hub genes were overexpressed in conjunction with KRAS in the heatmap of cohorts of KRAS-mutant lung cancer patients. Furthermore, GSEA analysis revealed that the cell cycle route had the highest significance, with a p-value of 3.498e-08 (Fig. 8). KRAS thereby upregulates the genes associated with the cell cycle subtype. Collectively, these hub genes might participate in the advancement of NSCLC induced by EGFR or KRAS mutations.

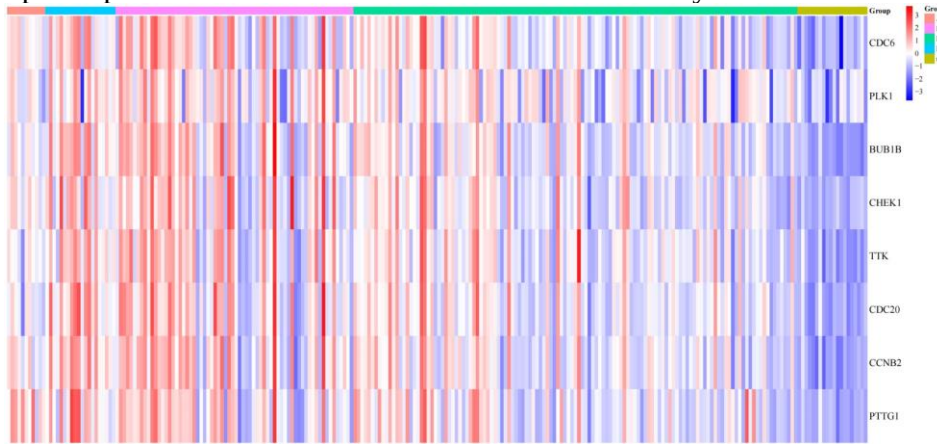


Fig. 7 Heatmap showing manifestation of the center gene in lung cancers with and without KRAS mutations in comparison to normal lung tissues (GSE31210)

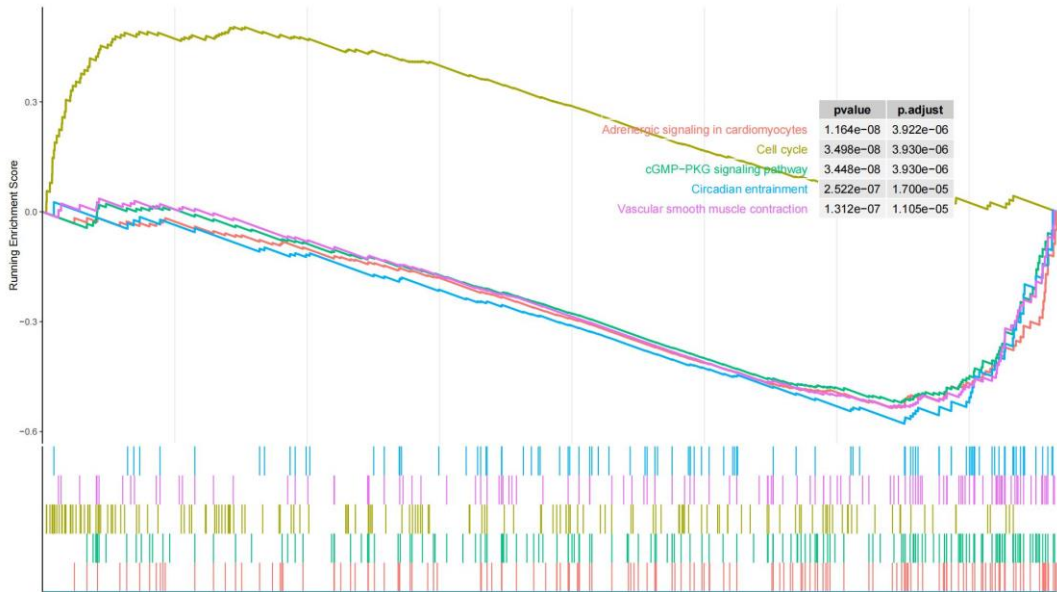


Fig. 8 Using gene set enrichment analysis (GSEA), the top five paths for enrichment in NSCLC samples with elevated KRAS expression were found

3.4. Verification of Gene Expression Associated with the Cell Cycle in NSCLC Tissues and Healthy Tissues at the mRNA and Protein Levels

3.4.1 Hub Gene mRNA expression levels rose in NSCLC

With the goal to confirm the hub gene expression levels in healthy and NSCLC tissues, We made use of the Gene Expression Profiling Interactive Analysis (GEPIA) (<http://gepia.cancer-pku.cn/>). In comparison to normal tissues, we discovered that the hub genes' degrees of expression (CDC6, PLK1, BUB1B, CHEK1, TTK, CDC20, CCNB2, and PTTG1) were markedly up-regulated in NSCLC tissues (Fig. 9). It is worth noting that the degree of hub gene expression were marginally elevated in LUSC compared to LUAD.

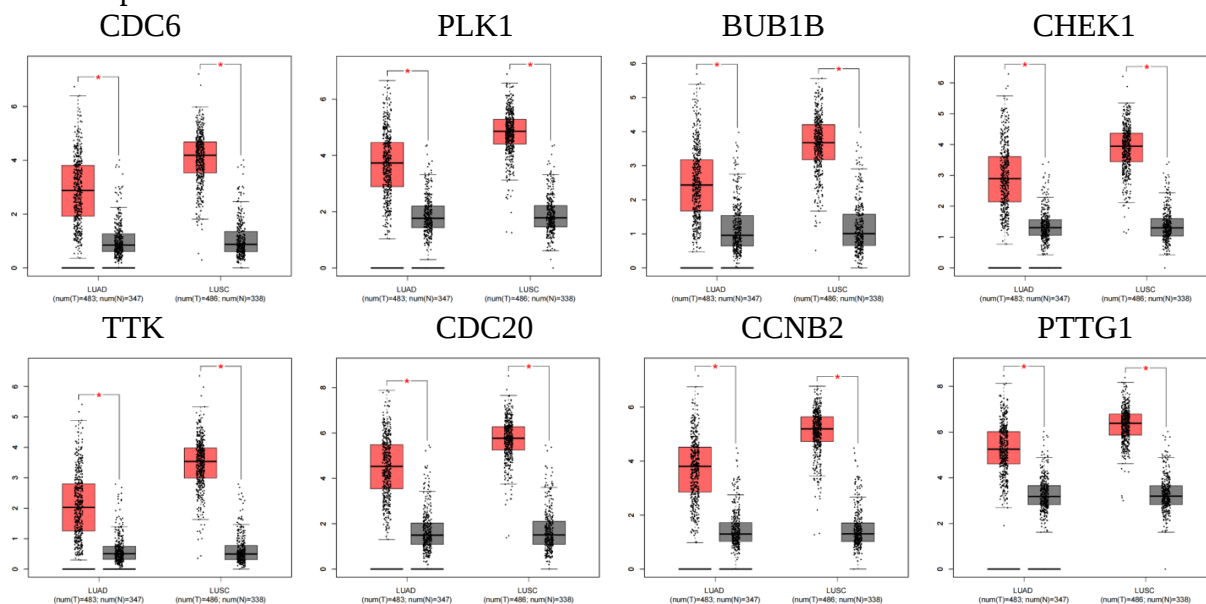


Fig. 9 Boxplots according to the TCGA that display the hub gene expression levels (grey box represents normal tissue, red box represents tumour tissue). The boxes in red and grey respectively correspond to NSCLC healthy tissues

3.4.2 Hub Gene Protein Expression Levels Rose in NSCLC

Furthermore to examining the manifestation of messenger RNAs (mRNAs), We carried out a thorough examination of the amounts of protein expression associated with the core genes. The HPA (<http://www.proteinatlas.org/>) contains CDC6, PLK1, TTK, CDC20, CCNB2, and PTTG1 protein immunohistochemistry data. Based on the data obtained from the database, it can be shown that a majority of the hub genes had either undetectable levels or displayed decreased lung tissue expression, as depicted in Fig. 10.

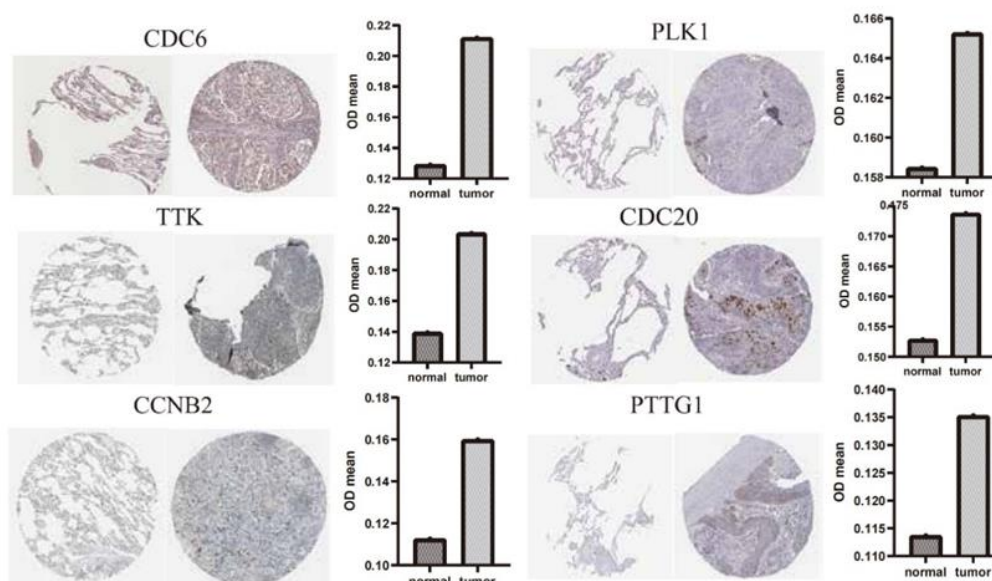


Fig. 10 Hub gene immunohistochemistry malignant cells from NSCLC that are included in the HPA database

3.5. An Analysis of Hub Gene Regulation Mechanisms Downstream

3.5.1 Amplification of Hub Genes Contribute to TKI Resistance in NSCLC Caused by EGFR Mutations

Lung cancer is mostly caused by mutations in the EGFR gene. Hub gene and EGFR correlation scatterplots were produced by using the Correlation module of GEPIA to analyse the LUAD, LUSC, and lung tumours (Fig. 11). The graphic demonstrated a link between EGFR and hub genes. Furthermore, compared to EGFR TKI-sensitive tissues from lung cancer, individuals with EGFR TKI-resistant tissues from lung cancer (GSE161584) had substantially greater degrees of manifestation of the key genes (PLK1, BUB1B, CHEK1, TTK, CDC20, CCNB2, and PTTG1) (Fig. 12). In conclusion, hub gene overexpression may be a factor in TKI resistance in EGFR-mutated NSCLC.

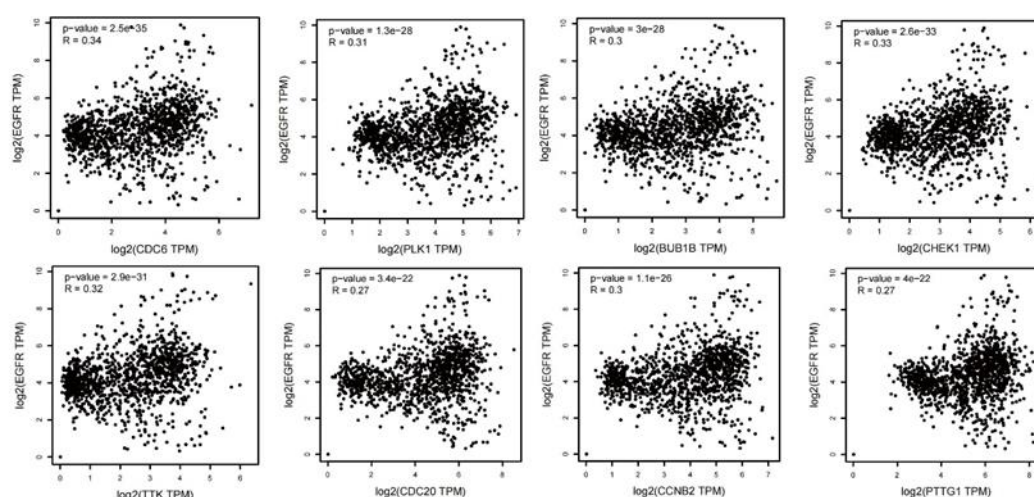


Fig. 11 Scatterplots showing the connections between EGFR and hub gene expression in the TCGA cohort

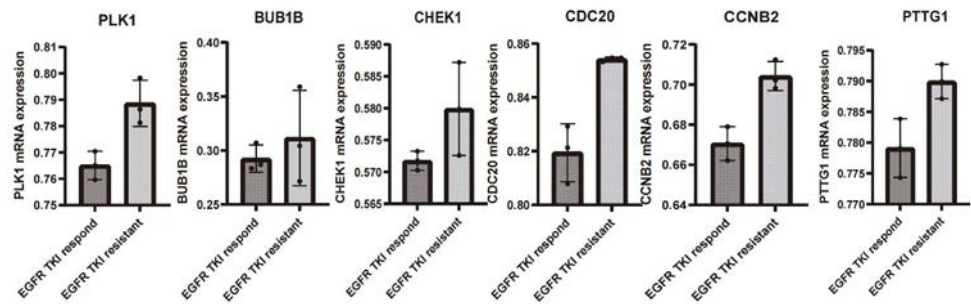


Fig. 12 Hub gene expression in lung cancer tissues that are TKI-resistant and EGFR-TKI-sensitive from the GSE161584 dataset

3.5.2 Resistance in NSCLC by EGFR Mutations due to Gene Copy Number Alteration Frequencies

Hub genes may contribute to resistance in NSCLC brought on by EGFR mutations by altering the frequencies of gene copy number alteration (CAN). Subsequently, our investigation aimed to evaluate the levels of gene expression for the central genes by using publicly available cancer genomics datasets on cBioPortal (<http://www.cbioportal.org/>). In our study, it was shown that the genomes of 2,976 patients from 8 studies on NSCLC had varying degrees of alterations in hub genes. Among the examined genes, the frequency of gene modifications in hub genes was found to be less than 1%. Furthermore, each gene exhibited distinct CNA patterns. The bulk of important genes typically underwent copy gain or shallow depletion, while CAN events in deep depletion and amplification occurred less frequently. (Fig. 13). In addition, we also found that the alterations of BUB1B and PLK1, PLK1 and TTK, CHEK1, and TTK had the tendency of co-occurrence (Table 1).

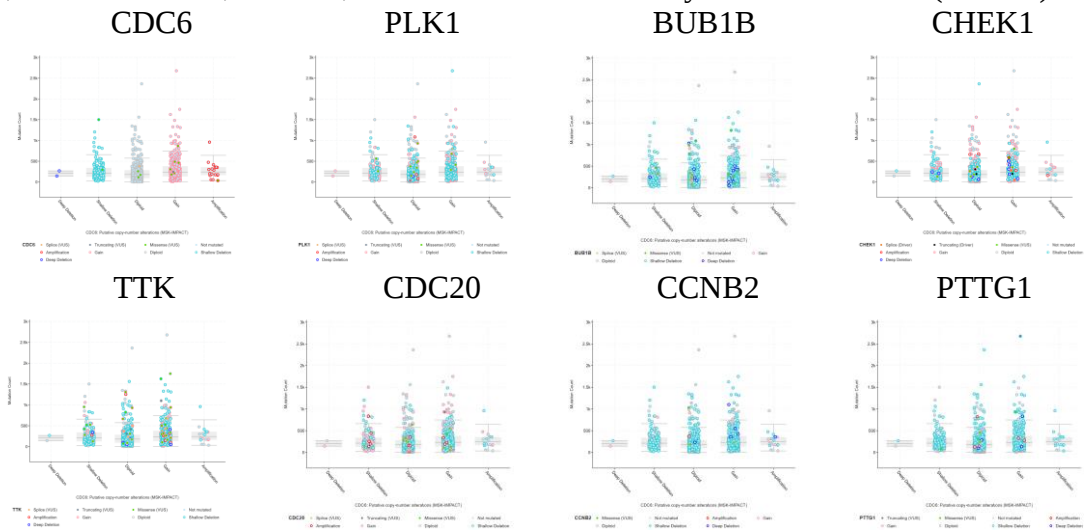


Fig. 13 Less than 1% of hub genes had altered genes, and each gene had a distinct CNA type

Table 1. In NSCLC, three core gene pairs are more likely to hold CNAs simultaneously

A	B	Neither	A to B	B to A	Both	Log2 Odds Ratio	p-Value	Tendency
BUB1B	CCNB2	1636	26	7	3	>3	0.001	Co-occurrence
PLK1	TTK	1603	20	56	3	2.102	0.044	Co-occurrence
CHEK1	TTK	1586	37	55	4	1.640	0.053	Co-occurrence

4. Discussion

4.1. Identification of Key Nodes

Lung cancer is widely accepted as a major contributor to high lethality owing to lacking related knowledge of hub genes. Therefore, identifying several genes that tip a scale can fill some parts of the blank in the arena, building on the work of several methods of data processing and pathway analysis, and the method [7] had been previously used to find out DEGs. First of all, by overlapping DEGs, we found eight genes as the figure showed CDC6, PLK1, BUB1B, CHEK1, TTK, CDC20, CCNB2, and PTTG1. This indicated that these eight genes might perform an essential function in the prognosis of patients. Then we did further research based on survival analysis, and the figure showcased that these genes all expressed lower in the patients who suffered from lung cancer and the lower expression of genes could be linked to higher survival rate. From this, we made a hypothesis whether these genes were significant nodes in certain pathway.

4.2. Potential Mechanism of Regulation in Cell Cycle Pathway

PPI network and GO pathway analysis were used to process the data from two databases and identified one enriched cell cycle pathway. The previous article has revealed that CDC20 contributed to the dormancy of cell and led to the EMT pathway [8], which was pivotal to the treatment of cancer. However, BubR1 expressed by BUB1B curbed the function of CDC20 [9], which could have a devastating effect on the prognosis of lung cancer. Combined with the expression of the gene, its lower expression might reduce the inhibition of CDC20, thus having a positive impact on the prognosis.

4.3. EGFR TKI Resistance is Related to the High Expression of Certain Genes

Moreover, we made an analysis of mRNA expression of these genes in two groups of patients. One could respond to EGFR TKI, and the other was resistant to it. Previous articles showed that a mutation in EGFR was found in a majority of patients who suffered from lung cancer [10], and the patients showed good prognosis after being treated by EGFR TKI medication. From the figure, we could identify that These genes' mRNA expression was greater in the resistance group, which was connected to a poorer prognosis.

4.4. One of the Flaws in the Experiment

In light of the fact that predecessors had done few investigations in the cell cycle pathway, the results above could verify the outcomes of analysis and provide an innovative way for future research in targeted therapy. Nevertheless, there still existed some deficits in our experiment. At first, we overlapped genes of LUSC and LUAD from TCGA database and the whole genome from GSE19188 in GEO database due to a lack of related data in the latter database. If we had overlapped the data from the same source but different databases, the scope would have been narrowed down and more precise.

Author Contributions

All the authors contributed equally and their names were listed in alphabetical order.

Chen, Meixi designed the research and drafted the manuscript; Chen, Meixi collected data and drafted the figures; Pan, Jiajing was responsible for the introduction section of the paper; Wang, Guanyang completed the discussion section; Wang, Guanyang and Pan, Jiajing critically revised the manuscript. All authors read and approved the final manuscript.

References

- [1] SIEGEL R L, MILLER K D, FUCHS H E, et al. Cancer Statistics, 2021. *CA: a cancer journal for clinicians*, 2021, 71(1): 7-33.
- [2] ARBOUR K C, RIELY G J. Systemic Therapy for Locally Advanced and Metastatic Non-Small Cell Lung Cancer: A Review. *Jama*, 2019, 322(8): 764-74.
- [3] CHEN Z, FILLMORE C M, HAMMERMAN P S, et al. Non-small-cell lung cancers: a heterogeneous set of diseases. *Nature reviews Cancer*, 2014, 14(8): 535-46.
- [4] WANG M, HERBST R S, BOSHOFF C. Toward personalized treatment approaches for non-small-cell lung cancer. *Nature medicine*, 2021, 27(8): 1345-56.
- [5] RICCIUTI B, ALESSI J V, ELKRIEF A, et al. Dissecting the clinicopathologic, genomic, and immunophenotypic correlates of KRAS(G12D)-mutated non-small-cell lung cancer. *Annals of oncology : official journal of the European Society for Medical Oncology*, 2022, 33(10): 1029-40.
- [6] MILLER M, HANNA N. Advances in systemic therapy for non-small cell lung cancer. *BMJ (Clinical research ed)*, 2021, 375: n2363.
- [7] TANG Z, KANG B, LI C, et al. GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic acids research*, 2019, 47(W1): W556-w60.
- [8] DAI Y, WANG L, TANG J, et al. Activation of anaphase-promoting complex by p53 induces a state of dormancy in cancer cells against chemotherapeutic stress. *Oncotarget*, 2016, 7(18): 25478-92.
- [9] TANG Z, BHARADWAJ R, LI B, et al. Mad2-Independent inhibition of APC^{Cdc20} by the mitotic checkpoint protein BubR1. *Developmental cell*, 2001, 1(2): 227-37.
- [10] WANG Y, SCHMID-BINDERT G, ZHOU C. Erlotinib in the treatment of advanced non-small cell lung cancer: an update for clinicians. *Therapeutic advances in medical oncology*, 2012, 4(1): 19-29.