International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:14, No:12, 2020

Transcriptomics Analysis on Comparing Non-Small Cell Lung Cancer versus Normal Lung, and Early Stage Compared versus Late-Stages of Non-Small Cell Lung Cancer

Achitphol Chookaew, Paramee Thongsukhsai, Patamarerk Engsontia, Narongwit Nakwan, Pritsana Raugrut

***Abstract*—**Lung cancer is one of the most common malignancies and primary cause of death due to cancer worldwide. Non-small cell lung cancer (NSCLC) is the main subtype in which majority of patients present with advanced-stage disease. Herein, we analyzed differentially expressed genes to find potential biomarkers for lung cancer diagnosis as well as prognostic markers. We used transcriptome data from our 2 NSCLC patients and public data (GSE81089) composing of 8 NSCLC and 10 normal lung tissues. Differentially expressed genes (DEGs) between NSCLC and normal tissue and between early-stage and late-stage NSCLC were analyzed by the DESeq2. Pairwise correlation was used to find the DEGs with false discovery rate (FDR) adjusted p-value ≤ 0.05 and |log2 fold change| ≥ 4 for NSCLC versus normal and FDR adjusted p-value ≤ 0.05 with |log2 fold change| ≥ 2 for early versus late-stage NSCLC. Bioinformatic tools were used for functional and pathway analysis. Moreover, the top ten genes in each comparison group were verified the expression and survival analysis via GEPIA. We found 150 up-regulated and 45 down-regulated genes in NSCLC compared to normal tissues. Many immnunoglobulin-related genes e.g., IGHV4-4, IGHV5-10-1, IGHV4-31, IGHV4-61, and IGHV1-69D were significantly up-regulated. 22 genes were up-regulated, and five genes were down-regulated in late-stage compared to early-stage NSCLC. The top five DEGs genes were KRT6B, SPRR1A, KRT13, KRT6A and KRT5. Keratin 6B (KRT6B) was the most significantly increased gene in the late-stage NSCLC. From GEPIA analysis, we concluded that IGHV4-31 and IGKV1-9 might be used as diagnostic biomarkers, while KRT6B and KRT6A might be used as prognostic biomarkers. However, further clinical validation is needed.

***Keywords*—**Bioinformatics, differentially expressed genes, non-small cell lung cancer, transcriptomics.

I. Introduction

LUNG cancer is uncontrollable cell growth that forms in the tissues of the lung, usually in the epithelial cells in lining air passages [1]. In 2018, there were 2.1 million new lung cancer cases and 1.8 million deaths, respectively. Two main histologic groups are recognized: small cell lung cancer and non-small cell lung cancer [2]. Approximately 85% of lung cancers are NSCLC [3]. Because of the limitations of chest X-ray, up to 25% of all patients do not reveal the malignant tumor [4]. Over 40% of primary NSCLC patients are already diagnosed with advanced-stage cancer which leads to with poor survival outcome.

RNA sequencing is a recent approach for transcriptome profiling, which investigates a set of all RNA in the cell presented in any condition [5]. Profiling of gene expression using RNA-sequencing provides a high-resolution view of the transcript.

Identifying biomolecules in tumor tissues is a valuable source of biological samples because they carry original information [6]. However, the publications regarding gene expression between early-stage and late-stage NSCLC were limited. Herein, we identify the DEGs between NSCLC and normal lung, also early stage compared with late-stage NSCLC. The DEGs identified in this study will be usual for further research as candidate markers for diagnostic and prognostic biomarkers in NSCLC.

A. Chookaew is with the Department of Biomedical Sciences, Faculty of Medicine, Prince of Songkla University, Songkhla, 90110 Thailand (phone: +66918476577; e-mail: achitphol39@hotmail.com).

P. Thongsukhsai is with the Department of Pathology, Faculty of Medicine, Prince of Songkla University, Songkhla, 90110 Thailand (e-mail: tparamee@gmail.com).

N. Nakwan is with the Hat Yai Hospital, Songkhla, 90110 Thailand (e-mail: naronak@hotmail.com).

P. Engsontia is with the Division of Biological Science, Faculty of Science, Prince of Songkla University, Songkhla, 90110 Thailand (e-mail: patamarerk.e@psu.ac.th).

P. Raugrut is with the Department of Biomedical Sciences, Faculty of Medicine, Prince of Songkla University, Songkhla, 90110 Thailand (e-mail: rpritsan@ medicine.psu.ac.th).

II. Material and Method

Tissue Samples and GEO Data Set

We used transcriptome data from 2 NSCLC samples, collected from Hat Yai hospital, Sonkhla, combined with data from 8 samples of public data set (GSE81089) from the GEO database. We used the fasterq-dump package for retrieving those data. The total 10 samples were 5 early-stage and 5 late-stage NSCLCs. Both early and late-stage NSCLC groups consisted of 1 ADC and 4 SCC subtypes. Moreover, 10 normal lung transcriptome data were also downloaded from the same dataset.

Our RNA samples were extracted from the two patients' tumor tissues using the RNeasy Mini kit (QIAGEN, Germany) according to the manufacturer's protocol [7]. All procedures were performed at 4 ℃ in an RNase-free environment.

RNA Sequencing

Our samples were sequenced by Macrogen Incorporation. A total of 1 ug RNA per sample was used for RNA sample

preparations. TruSeq library construction was then generated. The Illumina NovaSeq 6000 platform was used for 150 paired-end sequencings in our 2 NSCLC samples. However, 100 bp paired reads were generated in the GSE81089 dataset with an Illumina HiSeq 2500.

Data Processing and DEGs

Transcripts from RNA sequencing were formatted in FASTQ. Erroneous sequence variants were introduced due to the library preparation step or even the sequencing step. Therefore, the qualities of these sequence data were assessed [8]. We used FastQC and Trimmomatic packages for checking the quality and trimming the sequence, respectively. The trimmed sequences were aligned with the human genome reference (Genome Reference Consortium GRCh38) using STAR. Then, abundances of the sequenced reads that can map to the human genome reference were quantified. After that, the number of reads was counted by gene using the RSEM. The FPKM was used for normalization. DEGs were analyzed by the DESeq2. The DEGs were regarded as the candidate genes and used for downstream analysis, functional analysis.

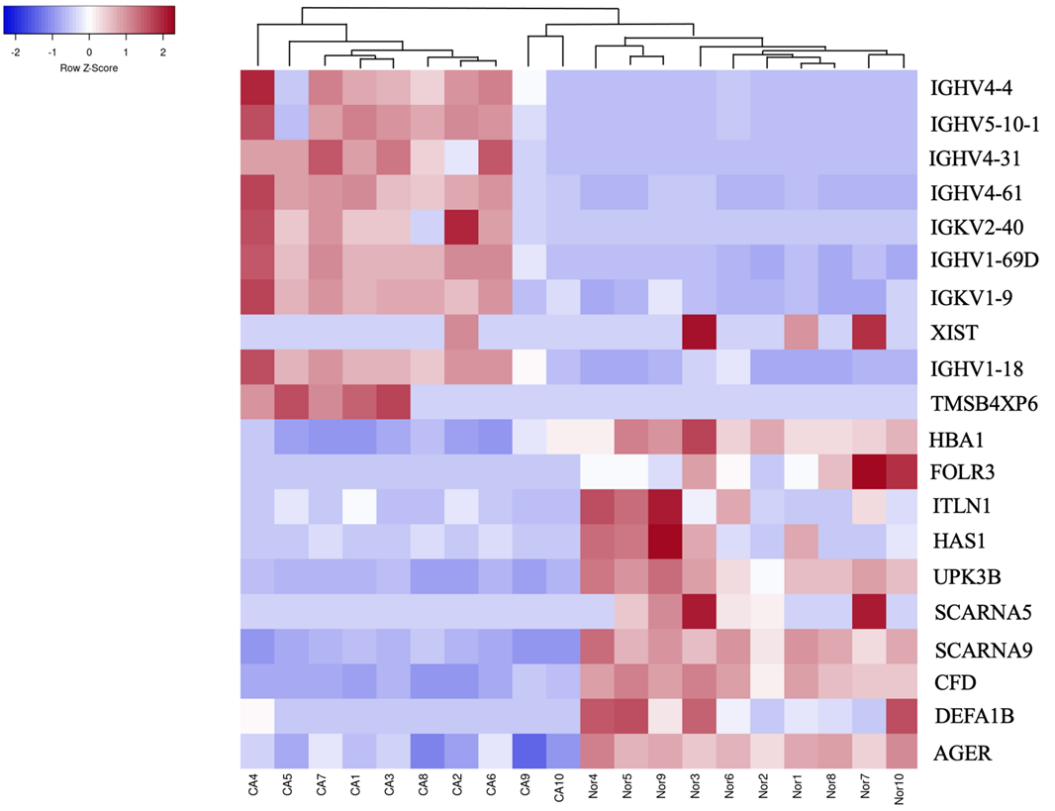
Statistical Analysis

The pairwise correlation was used to find DEGs. The adjusted p-values were applied using Benjamini and Hochberg FDR method. The FDR adjusted p-value ≤ 0.05, and $|\log_2$ fold change$| \geq 4$, were considered significant in NSCLC versus normal and the comparison group while the FDR adjusted p-value ≤ 0.05, and $|\log_2$ fold change$| \geq 2$ were considered significant for early versus late cancer comparison group.

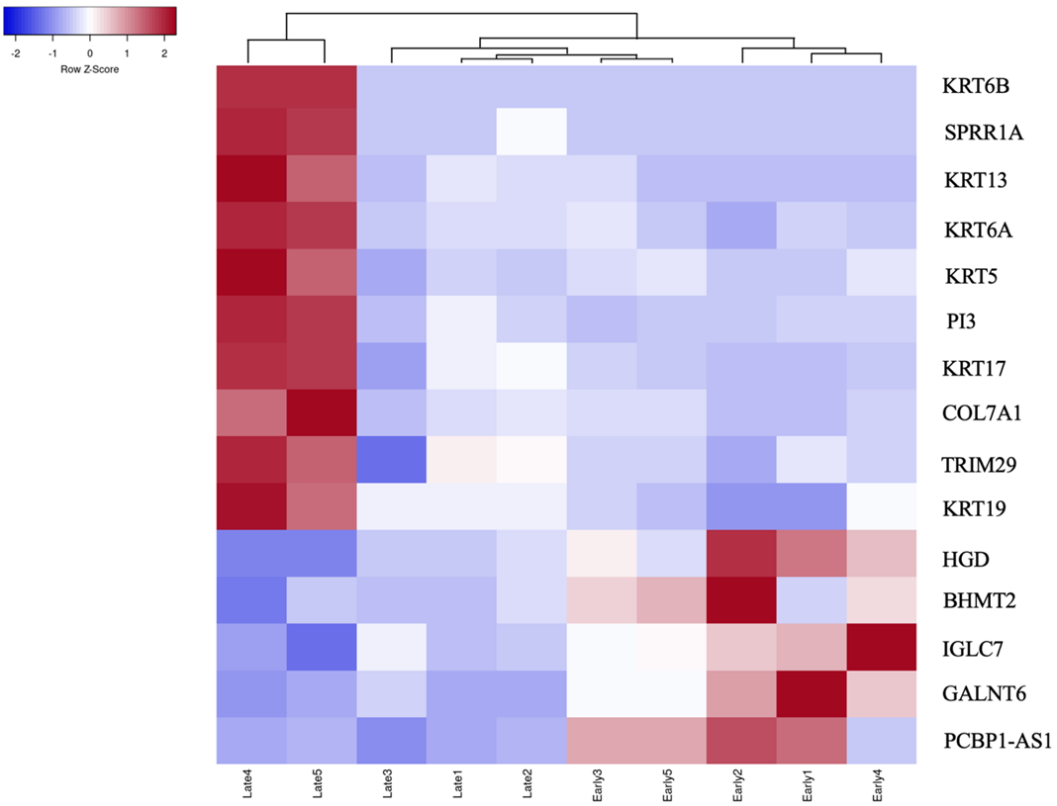
III. Results

Identification of DEGs

Comparing NSCLC with normal lung transcriptomes provided 150 up-regulated and 45 down-regulated DEGs. We also found that the expression of 27 genes was significantly different in late versus early stage; among these 22 genes were up-regulated, and 5 genes were down-regulated in late-stage compared to early-stage cancer. Representatives of top up-regulated and down-regulated genes by fold change in each comparison group were shown in Tables I and II. Heat map plotted by the heatmapper (http://www.heatmapper.ca/expression/) indicating the DEGs on both comparison groups is shown in Figs. 1 (a) and (b).



(a)



(b)

Fig. 1 (a) Heat map of gene expression on NSCLC compared with normal lung tissues; (b) and on early stage compared with late stage NSCLC tissues. Nor, normal; CA, cancer.

As shown in Table I, among the top 10 DEGs between NSCLC and normal lung tissues, IGHV4-4 and HBA1 were markedly increased and decreased, respectively, whereas KRT6B and HGD were markedly increased and decreased, respectively between early- and late-stage NSCLC (Table II).

TABLE I
DEGs BY RNA-SEQ PROFILING IN NSCLC COMPARED WITH NORMAL LUNG TISSUES

Gene symbol	Log_2 FC	p-value	Adj p-value
IGHV4-4	11.81504	9.07E-15	1.02E-12
IGHV5-10-1	9.365471	2.37E-11	1.43E-09
IGHV4-31	9.341548	1.45E-15	1.93E-13
IGHV4-61	9.100569	4.21E-20	1.91E-17
IGHV1-69D	8.787101	3.93E-25	3.48E-22
IGKV1-9	8.770373	2.08E-19	7.94E-17
XIST	8.552698	2.19E-08	6.13E-07
HBA1	-7.221870	1.60E-15	2.08E-13
FOLR3	-6.239054	7.60E-08	1.80E-06
ITLN1	-6.166310	1.98E-07	4.01E-06

Gene Ontology Analysis of DEGs

We identified the biological signatures of DEGs by using PANTHER (http://pantherdb.org/). We found that complement activation, regulation of leukocyte activation, membrane disruption in other organisms, immune response-activating signal transduction, immunoglobulin receptor binding, antigen binding, immunoglobulin complex, immunoglobulin complex, circulating, and extracellular region, were significantly decreased on DEGs in NSCLC compared with human normal lung tissues. In contrast, epidermis development, cornification, keratinization, skin development, keratinocyte differentiation, structural constituent of the cytoskeleton, structural molecule activity, intermediate filament cytoskeleton, extracellular space, and intermediate filament were significantly increased on DEGs in early-stage NSCLC compared with late-stage NSCLC tissues (Figs. 2 and 3).

TABLE II
DEGs BY RNA-SEQ PROFILING IN LATE-STAGE COMPARED WITH EARLY-STAGE NSCLC

Gene symbol	Log_2 FC	p-value	Adj p-value
KRT6B	22.679395	6.78E-14	7.33E-10
SPRR1A	22.275281	1.84E-13	9.94E-10
KRT13	12.266376	1.28E-05	9.22E-03
KRT6A	9.1123810	7.02E-07	9.47E-04
KRT5	9.0394090	4.38E-06	3.94E-03
PI3	8.073885	9.87E-09	2.66E-05
KRT17	7.449737	8.63E-09	2.66E-05
COL7A1	6.154398	3.53E-06	3.47E-03
HGD	-4.609089	7.38E-05	0.030635131
BHMT2	-4.334598	1.34E-04	0.045091792

Pathway Enrichment Analysis

We investigated the biological signatures of these DEGs in both groups of comparison Reactome (https://reactome.org/). We found that the DEGs of NSCLC compared with normal lung mostly converged on processes involving immune response and hemostasis. The DEGs of late-stage NSCLC compared with early-stage NSCLC had close association with cornification formation, keratinization, and cell development (Figs. 4 (A) and (B)).

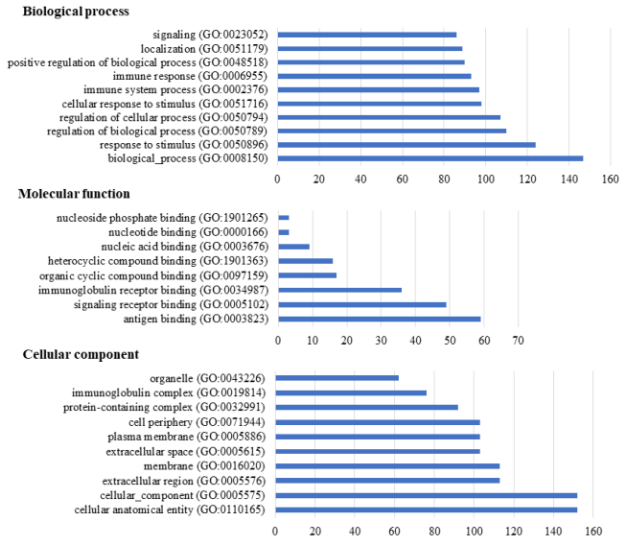


Fig. 2 PANTHER results for gene ontology on DEGs in NSCLC compared with normal lung

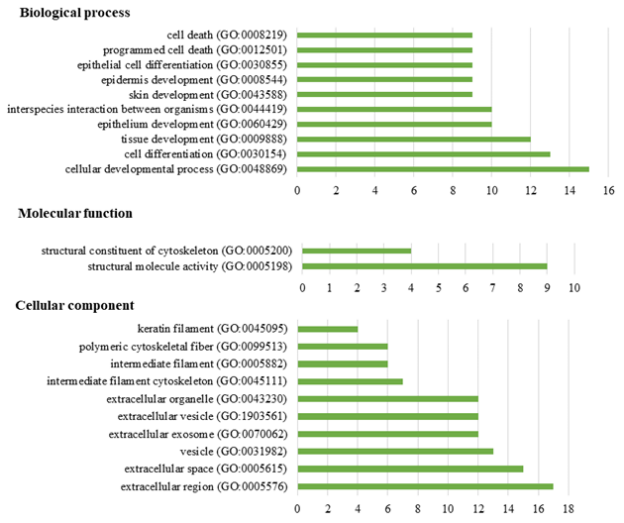


Fig. 3 PANTHER results for gene ontology on DEGs in late-stage NSCLC compared with early-stage NSCLC

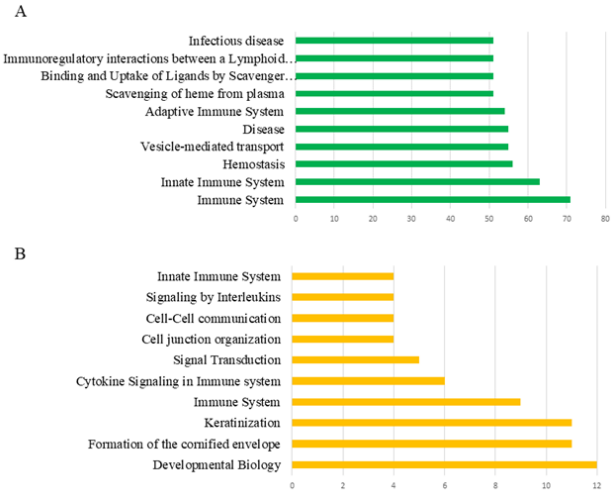


Fig. 4 Biological annotation for DEGs of NSCLC. (A) Biological annotation for DEGs on comparing NSCLC with normal lung; (B) Biological annotation for DEGs on comparing late stage with early-stage NSCLC

Verification of Top DEGs through GEPIA

We verified our top ten DEGs of each comparison group using a larger data set through GEPIA (http://gepia.cancer-pku.cn/index.html), which is an interactive web server for analyzing the RNA sequencing expression data of various normal and tumor samples from the TCGA and the GTEx projects. The verification results confirmed the significant differential expression between normal and NSCLC samples. We set the cutoff to plot the box plot as $|\log2FC| \geq 2$ and p-value ≤ 0.05. As shown in Table I, only the 5 genes, IGHV4-31, IGHV1-69D, IGKV1-9, IGHV1-18, and HBA1, were significant via GEPIA in both adenocarcinoma (ADC) and squamous cell carcinoma (SCC) compared with normal lung. All of these genes were not significantly correlated with overall survival in patient. The top 2 significantly expressed genes (IGHV4-31 and IGKV1-9) of both ADC and SCC subtypes compared with normal lung were shown in Figs. 5 (A) and (B). Also, some of the DEGs from early and late-stage NSCLC were significantly correlated with survival in only ADC, including KRT6B, KRT6A, KRT17, and COL7A1 genes. Two examples of significant genes related to survival were shown in Figs. 5 (C) and (D).

IV. Discussion

In this study, we identified genes and their functions that show expression correlated to NSCLC (A total of 195 genes: 150 up-regulated and 45 down-regulated DEGs in NSCLC versus normal tissues). IGHV4-4, one of the high expression genes in this group, is an immunoglobulin heavy variable 4-4 that plays an important role in antigen recognition [9]. Most DEGs in this group were related to the immune response, such as IGHV4-4, IGHV5-10-1, IGHV4-31, and IGHV4-61. All these genes play an essential role in immunoglobulin binding and antigen-binding [9]. Our results support the notion that cancer development may activate host immune function.

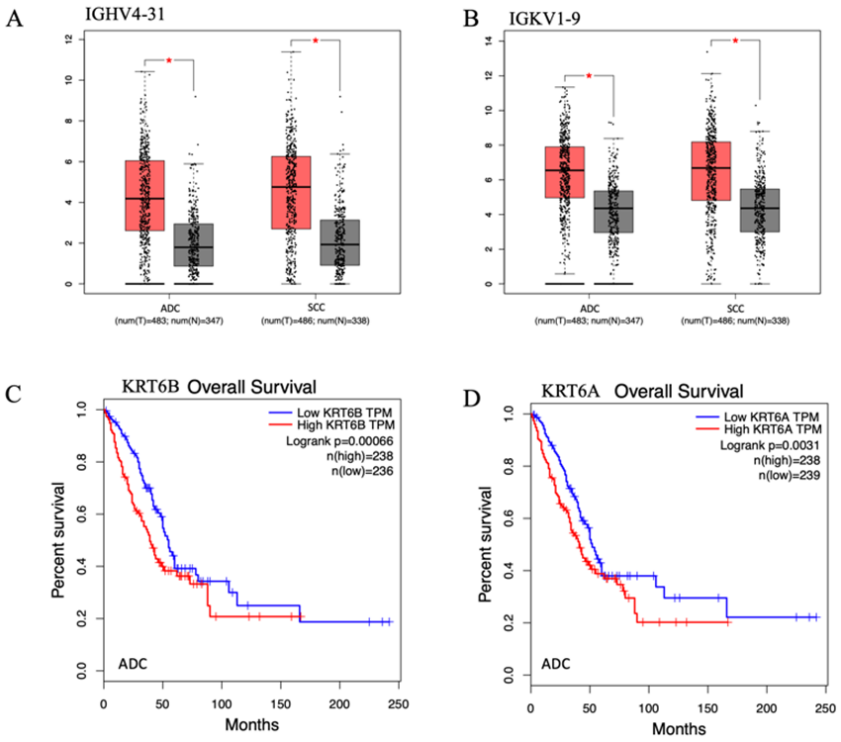


Fig. 5 (A) and (B) Box plots demonstrated significant up-regulation of IGHV4-31 and IGKV1-9 in ADC and SCC (red) compared to normal tissue (black). (C) and (D) High expression of KRT6B and KRT6A were associated with poorer survival in ADC ($P \leq 0.05$)

For early versus late-stage NSCLC comparison, 27 DEGs were identified. Many genes are related to cell development and cancer progression. KRT6B, which is one of the highest expressed genes in this comparison group, is keratin 6B [10]. Moreover, various KTR members, such as KRT13, KRT6A, KRT5, and KRT17 were also significantly increased. KRT17 markedly promoted tumor growth and invasion. High expression of KRT17 was significantly correlated with poor overall survival in ADC [11]. KRT6A, KRT6B, and KRT6C were reported as potential markers for distinguishing between ADC and SSC [12]. However, we found that high expression of KRT6B was significantly correlated with poor overall survival in ADC but not significant in SCC. Our finding provides evidence that some genes identified from DGE analysis are associated with cancer development and progression, and some genes can be used as potential prognostic biomarkers. However, this study used only 10 samples in both NSCLC and normal lung tissue. Analysis using a larger sample size is needed.

V. Conclusion

Our study found that various immune-related genes, such as IGHV4-31 and IGKV1-9 may play a role in cancer development. They may be used as potential diagnostic biomarkers for NSCLC. KRT6B and KRT6A were up-regulated in late-stage NSCLC, which may be potential prognostic biomarkers in ADC. The results, however, have to be validated in real clinical samples.

Acknowledgment

This study was funded by the Faculty of Medicine, Prince of Songkla University, Thailand. Moreover, the authors thank for sharing the data from The Gene Expression Omnibus (GEO) database.

References

[1] R. Siegel, D. Naishadham, A. Jemal, "Cancer statistics, 2013" *CA Cancer J Clin*, vol. 63, no.1, pp. 11-30, Jan. 2013.
[2] F. Bray, J. Ferlay, I. Soerjomataram, R. Siegel, L. Torre, A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries" *CA Cancer J Clin*, vol. 68, no. 6, pp. 394-424, Jul. 2018.
[3] S. Ramalingam, C. Belani, "Systemic Chemotherapy for Advanced Non-Small Cell Lung Cancer: Recent Advances and Future Directions" *The Oncologist*, vol. 13, no. S1, pp. 5-13, Jan. 2008.
[4] Z. Fan, W. Xue, L. Li, C. Zhang, J. Lu, Y. Zhai, "Identification of an early diagnostic biomarker of lung adenocarcinoma based on co-expression similarity and construction of a diagnostic model" *J Transl Med*, vol. 16, no. 1, Jul. 2018.
[5] Z. Wang, M, Gerstein, M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics" *Nat. Rev. Genet*, vol. 10, no. 1, pp. 57-63, Jan. 2009.
[6] D. Carvajal-Hausdorf, K. Schalper, V. Neumeister, "Quantitative measurement of cancer tissue biomarkers in the lab and in the clinic" *Lab Invest*, vol. 95, pp. 385-396, Apr. 2015.
[7] Quigen. Total RNA purification from animal tissue, in: RNeasy® Mini Handbook. *Elsevier*, 2009, pp. 45-52.
[8] S. Pfeifer, "From next-generation resequencing reads to a high-quality variant data set" *Heredity*, vol. 118, pp. 111-124, Oct. 2017.
[9] M. Lefranc, "Immunoglobulin and T Cell Receptor Genes: IMGT(®) and the Birth and Rise of Immunoinformatics" *Front Immunol*, Vol. 5, no. 5, pp. 22, Feb. 2014.
[10] J. Xiao, X. Lu, X. Chen, Y. Zou, A. Liu, W. Li, B. He, S. He, Q. Chen, "Eight potential biomarkers for distinguishing between lung adenocarcinoma and squamous cell carcinoma" *Oncotarget*, vol. 3, no.

8, pp. 71759-71771, May. 2017.
[11] J. Liu, L. Liu, L. Cao, Q. Wen, "Keratin 17 Promotes Lung Adenocarcinoma Progression by Enhancing Cell Proliferation and Invasion" *Med Sci Monit*, vol. 24, pp. 4782-4790, Jul. 2018.
[12] J. Xiao, X. Lu, X. Chen, et al. "Eight potential biomarkers for distinguishing between lung adenocarcinoma and squamous cell carcinoma" *Oncotarget*, vol. 8, no. 42, pp. 71759-71771, May. 2017.