Expression pattern detection and classification of two types of lung cancer, LUAD and LUSC

Xin Wang

Abstract

Lung cancer is one of the most common cancers with a relatively high mortality rate, where lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) are the two most common lung cancer subtypes. Despite both being classified as non-small cell lung cancer (NSCLC), they are different in pathogenesis, prevalence, and prognosis, and thus should be treated differently under the goal of precision medicine. At the same time, machine learning shows great potential in biomedical-related tasks, especially in cancer type classification. How well can these methods be integrated and what interesting insights can they provide about LUAD and LUSC is the driving question of this project. To answer it, we first applied Differential Gene Expression Analysis (DGE) on a subset of the LUAD and LUSC datasets, respectively, where identified features were used to perform pathway enrichment analysis. Meanwhile, these features, together with the rest of the samples served as the dataset for building three classifiers, Logistic regression, Random forest, and XGBoost. Three classifiers all had good performance because of the clear expression patterns detected, especially for the Logistic regression model, which performed best with an F1 score higher than 0.95. Furthermore, explorations on feature importance suggested some features with large weights have biological significance, which might be the reason why logistic regression performed so well. And we also found that feature importance might serve as a tool for finding biomarkers.

1 Introduction

Cancer is the second leading cause of death worldwide, among which lung cancer, with a high mortality rate and low 5-year relative survival rate, is the leading cause of cancer-associated mortality[1]. More than 85% of lung cancer cases are classified as non-small cell lung cancer (NSCLC), which mainly contains two subtypes, lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC)[2]. Although LUAD and LUSC are NSCLC, they do differ in many aspects, for example, LUAD has a higher metastatic rate[3] while LUSC is more closely associated with smoking, and affects male more than female[4]. And one study believes that we should abandon the notion of NSCLC to develop more effective therapeutic procedures because LUAD and LUSC appear to be vastly distinct diseases at the molecular, pathological, and clinical level[5].

Performing Differential Gene Expression Analysis (DGE) on RNA-seq dataset is a standard method in silico for identifying possible biomarkers that can aid diagnosis and therapy, and further downstream analysis reveals which biological pathways the differentially expressed genes are enriched in. Meanwhile, with the advancement of machine learning and its wide application in a variety of domains, a number of classifiers for predicting cancer types based on RNA-seq data have been developed. By combining differential expression, pathway enrichment, and classification of LUAD and LUSC, we may be able to gain a new viewpoint on understanding them.

In this project, we downloaded both raw count and FPKM data of LUAD and LUSC from The Cancer Genome Atlas (TCGA)[6] and conducted a series of analyses. The differentially expressed features obtained through DGE on LUAD and LUSC datasets separately were not only used for Gene Ontology (GO)[7] enrichment analysis, but also treated as features for classification after taking the intersection. Then, three machine models, Logistic regression, Random forest, and Extreme Gradient Boosting (XGBoost) were built to classify LUAD, LUSC, and normal samples. Finally, the good performance of Logistic regression classifier was demonstrated by the following analyses of feature importance.

2 Study design and overview

To begin with, we downloaded the raw count and FPKM RNA-seq data of LUAD and LUSC from TCGA. The raw count data was divided into two subsets for both cancer types, one subset was then used to conduct differential expression analysis and GO enrichment analysis (Figure 1). The intersection of differentially expressed features in LUAD and LUSC, along with FPKM data of samples in the other subset were used for building three classifiers, Logistic regression, Random forest, and XGBoost. Then the model was evaluated using different performance metrics. Finally, we conducted feature importance exploration through feature weights in the Logistic regression model.

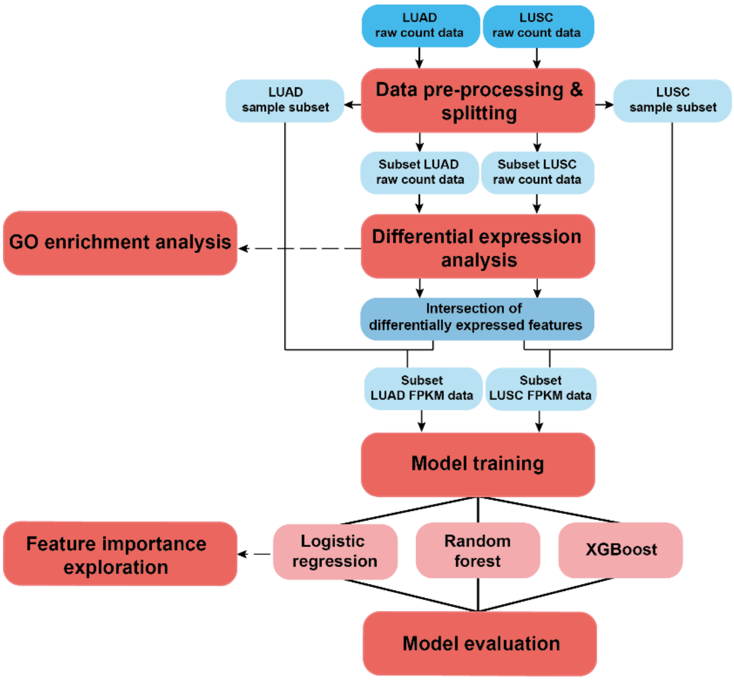

Figure 1. Project workflow.

3 Data pre-processing

There are 585 samples in the LUAD dataset and 550 samples in the LUSC dataset, both contain 60,489 features. Samples from Formalin-fixed paraffin-embedded (FFPE) tissue were eliminated to improve the precision of the study because FFPE tissue processing and sample storage have been suggested to substantially degrade RNA[8]. Then features with total counts less than the number of normal samples were removed to improve the efficiency of differential expression analysis.

To avoid overfitting, two raw count datasets were divided into two subsets, one subset was for the expression pattern detection, and the other subset was for classification. The criterion for the division was to have as similar LUAD and LUSC samples and tumor to normal ratio in classification task as possible (Table 1).

Table 1. Sample distribution after splitting

T: tumor sample, N: normal sample

Dataset	Pattern detection	Classification
LUAD	124T, 16N	388T, 42N
LUSC	111T, 14N	385T, 35N

4 Expression pattern detection for LUAD and LUSC

4.1 DGE analysis

DGE was carried out separately for LUAD and LUSC using R package DESeq2[9], and features with absolute log2 FC > 2, adjusted p-value < 0.05 were reported. Furthermore, genes with Ensembl ID were converted to gene symbols for easier identification, whereas genes without gene symbols retained their Ensembl IDs.

DGE identified 1042 down-regulated and 3349 up-regulated features in LUAD, while the numbers for LUSC are 2480 and 4475, respectively. Clear patterns were detected (Figure 2), and many differentially expressed genes (DEGs) in this project are associated with the corresponding lung cancer type, which is consistent with the results from previous studies. For example, the *PTPRH* gene (log2 FC = 6.13, adjusted_p_value = 1.89E-37) that encodes receptor-type protein tyrosine phosphatase was one of the most significant up-regulated genes in LUAD dected in this project, and it was found to be overexpressed in LUAD and might have prognostic implications[10]. In addition, *NEK2* gene (log2 FC = 4.13 and adjusted_p_value = 1.12E-35 in LUAD, log2FC = 6.83 and adjusted_p_value = 6.8E-83 in LUSC), which was found to be overexpressed in both LUAD and LUSC, was identified as an effective tumor proliferation marker of poor prognosis for NSCLC patients that can also help with therapeutic intervention[11].

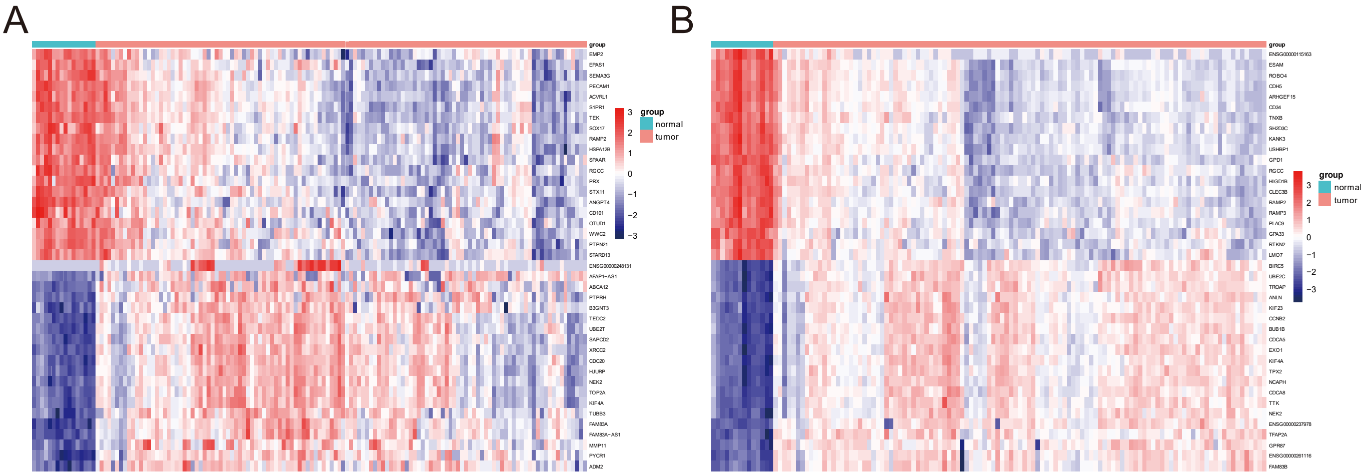

Figure 2. Heatmap of feature expression levels in (A) LUAD and (B) LUSC. Expression level of 20 most significant up-regulated and 20 down-regulated DEGs in LUAD and LUSC.

4.2 GO enrichment analysis

Go enrichment analysis was conducted through "*enrichGO*" function embedded R package clusterProfiler[12], with all three ontologies include Biological Process (BP), Molecular Function (MF), and Cellular Component (CC).

Results of GO enrichment analysis of DEGs showed differences between LUAD and LUSC at the biological pathway level (Figure 3). Most genes differentially expressed in LUAD were enriched in pathways related to the immune system, such as humoral immune response and defense response to the bacterium, while pathways related to cell development were enriched in LUSC. These observations are consistent with a study found that cell cycle promoting genes showed faster up-regulation in LUSC, whereas immune response promoting genes were more rapidly repressed in LUSC compared to LUAD[13]. Besides, keratinization-related pathways were also enriched in LUSC, which makes sense if most of the LUSC samples analyzed in this project belong to keratinizing squamous cell carcinomas (SCC), a subtype of SCC[14]. And keratinization of LUSC has been proved to be associated with poor clinical outcomes[15].

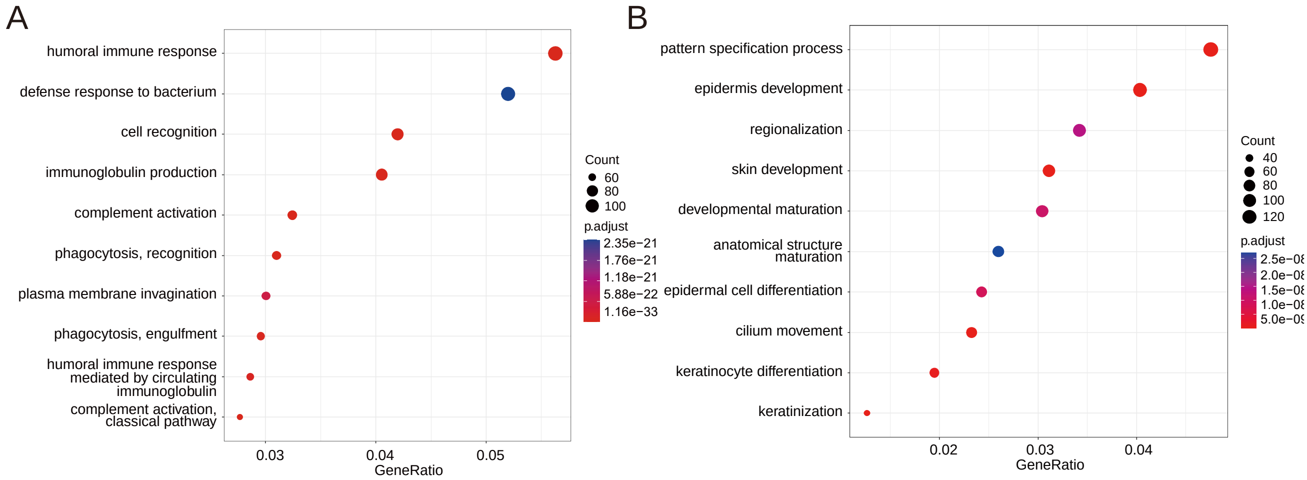

Figure 3. GO enrichment analysis on features differentially expressed in (A) LUAD and (B) LUSC. Top 10 pathways that the DEGs are enriched in (*ontology = "all"*).

5 Multi classification of LUAD, LUSC, and normal sample

The data for building classifiers came from a combination of samples in the second subset of raw LUAD and LUSC data, along with the intersection of features identified in LUAD and LUSC through DGE (2699 features). Normal samples from either LUAD or LUSC were labeled as "normal".

Before model construction, we first compared different procedures for handling two sets of DEGs by measuring their performance metrics. Although taking the union or difference of DEGs in LUAD and LUSC performed better, features with large weights failed to identify all three types of samples under the aforesaid circumstances, which can only be achieved by taking the intersection of two sets of DEGs. Models with log-transformed data performed better on average than models with the original data when it came to data scaling, hence log-transformation was applied to the data before model training.

With a 4:1 ratio of training and testing data, we developed three classifiers: Logistic regression, Random forest, and XGBoost. 5-fold cross-validation was used during training to get the optimal set of model parameters. In addition, *class weight = "balanced"* was set for each model to address the data imbalance issue. A micro score was employed to generate performance metrics in the evaluation section, which is a weighted score that can better reveal their performance without bias.

The Logistic regression model outperformed the other two classifiers, scoring the highest on all evaluation metrics. Furthermore, the majority of misclassifications for the Logistic regression model were between LUAD and LUSC, and all 15 normal samples were correctly classified (Figure 4). We believe there are two reasons for its success: 1) feature selection through DGE successfully selected features with great contributions to classification, enabling simple model to achieve relatively high scores; 2) the data became linearly separable after feature selection, making linear regression-based model the best among the three.

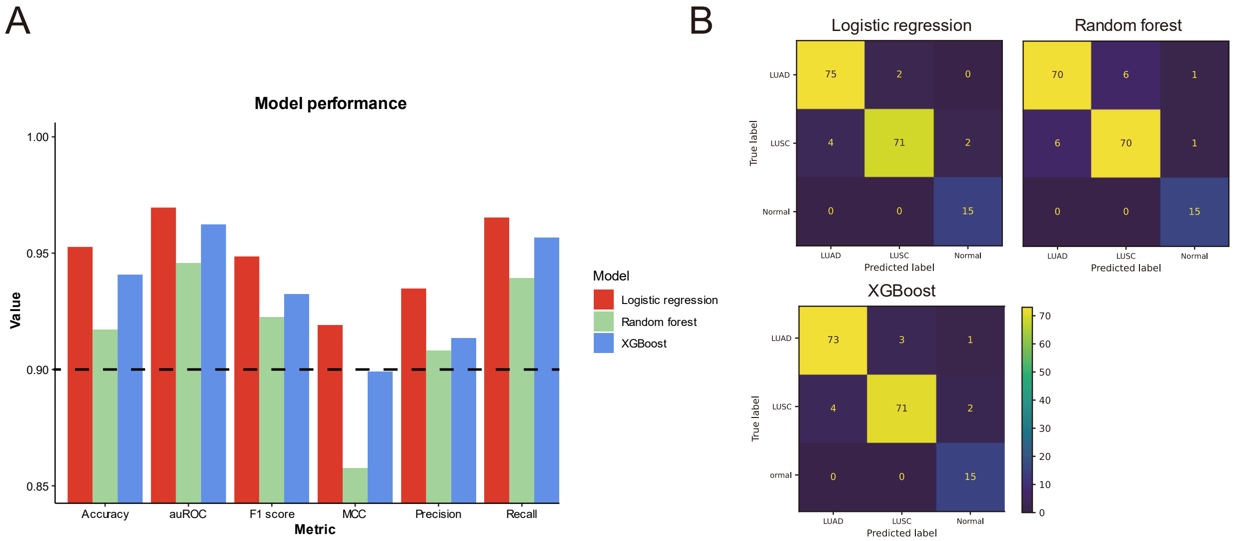

Figure 4. Performance of three classifiers. (A) Accuracy, auROC, F1 score, MCC, Precision and Recall of Logistic regression, Random forest, and XGBoost classifiers. (B) Confusion matrix of three classifiers

6 Feature importance exploration of Logistic regression model

In the Logistic regression classifier, we collected the weight of each feature for LUAD and LUSC classification, ordered absolute weights in descending order, and examined top features (Figure 5). Unsurprisingly, many features with large weights have been found in other studies to be associated with LUAD or LUSC. For example, *GCNT3* gene, which encodes a member of the N-acetylglucosaminyltransferase family, was found to be highly expressed in both NSCLC tissues, and its higher expression correlated significantly with advanced tumor-node-metastasis (TNM) stage[16]. Besides, *VWF* has been identified as a biomarker for LUAD[17], and its pseudogene *VWFP1* turned out to have a large weight in LUAD classification.

In terms of LUSC classification, *S100A7* and *KRT6A* caught our attention. *S100A7* encodes a calcium-binding protein and is highly expressed in lung cancer, especially in LUSC compared to LUAD[18]. And it also plays a role in biological pathways that are enriched by DEGs in LUSC, including skin development and epidermal cell differentiation. In addition, *S100A7* was found to negatively contribute to LUAD classification. The other gene *KRT6A* belongs to the keratin gene family. In LUAD, KRT6A was upregulated with increasing TNM stages, but its expression was significantly increased in advanced LUSC tumors[19], suggesting its stronger association with LUSC compared to LUAD.

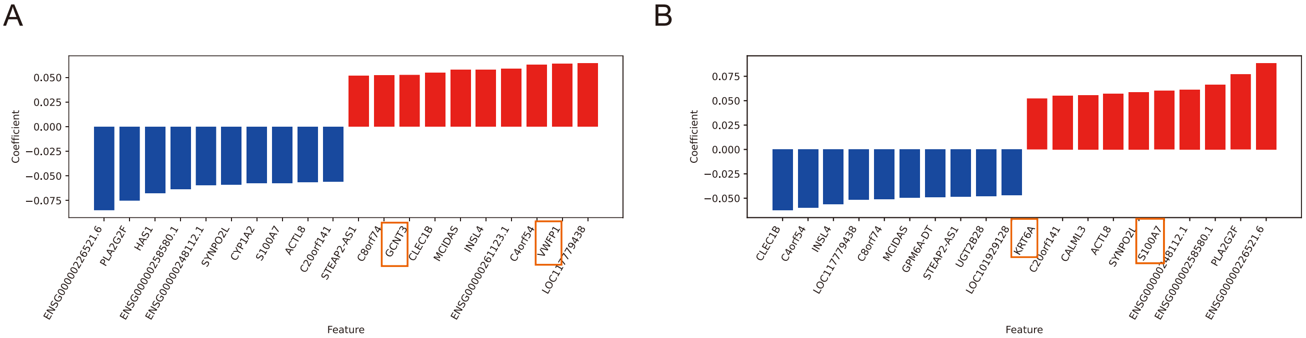

Figure 5. Features with top 10 positive weights and top 10 negative weights for LUAD class and LUSC class in Logistic regression classifier. (A) Top features and their weights for LUAD class. *GCNT3* and *VWFP1* were marked by an orange box. (B) Top features and their weights for LUSC class. *S100A7* and *KRT6A* were marked by an orange box.

Based on these findings, feature importance exploration might work as a tool to select potential biomarkers, and *S100A7* and *KRT6A* can be potential biomarkers for distinguishing between all three types of samples, LUAD, LUSC and normal sample.

7 Discussion

In this project, we focused on the differences between two lung cancer subtypes, LUAD and LUSC, in gene expression level, and conducted a series of analyses. DEGs and enriched biological pathways identified via DGE and GO enrichment analysis provided clear and different expression patterns for LUAD and LUSC. Most of the genes differentially expressed in LUAD were enriched in pathways related to immune system, whereas pathways related to cell development and keratinization-related pathways, were enriched in LUSC. Furthermore, three machine learning classifiers designed for multi-classification tasks all performed well, with the logistic regression classifier outperforming the others with all measures above or close to 0.95. When analyzing feature importance for the logistic regression classifier, the *S100A7* gene and *KRT6A* gene were found to have large weights in LUSC classification, and their overexpression in LUSC compared to LUAD was also observed in other studies, making them potential biomarkers for distinguishing LUAD and LUSC. This project integrated multiple commonly used bioinformatics methods for downstream analysis and machine learning methods to conduct a comprehensive analysis of LUAD and LUSC based on RNA-seq data and may provide insights on how can different methods be integrated.

There are several limitations of this project. First, both datasets were split into two parts for pattern detection and classification, leading to limited samples for classifier construction. With a small sample size, the robustness of models might be affected. Furthermore, we were unable to obtain an external dataset to evaluate the models we built because the external dataset did not include all features given into the models, making it unable to fit in the model. Lastly, this project can be more thorough if comparisons with models from other studies were included.

Although diagnosis of NSCLC is often through imaging tests including chest x-ray, computed tomography (CT) scan magnetic resonance imaging (MRI) scan and so on, in-silico research at gene expression level, such as this project, helps us pinpoint the affecting genes for each cancer type, reducing the search space for experiments at web lab, as well as finding out more risk factors such as tobacco smoking[20]. In terms of the treatment of NSCLC, although surgery, chemotherapy, and radiation therapy are still the predominant approaches, immunotherapy such as immune checkpoint blockade is developing rapidly[21,22]. A study shows that pembrolizumab monotherapy, which targets and blocks PD-1 on the surface of T-cells, provided durable antitumor activity and high 5-year OS rates in patients with advanced NSCLC [23], and identification of further possible neoantigens can help with the immunotherapy for NSCLC.

Reference

[1] Siegel, R.L., Miller, K.D. and Jemal, A. (2020), Cancer statistics, 2020. CA A Cancer J Clin, 70: 7-30. https://doi.org/10.3322/caac.21590

[2] Chen, Z., Fillmore, C. M., Hammerman, P. S., Kim, C. F., & Wong, K. K. (2014). Non-small-cell lung cancers: a heterogeneous set of diseases. Nature reviews. Cancer, 14(8), 535–546. https://doi.org/10.1038/nrc3775

[3] Milovanovic, I. S., Stjepanovic, M., & Mitrovic, D. (2017). Distribution patterns of the metastases of the lung carcinoma in relation to histological type of the primary tumor: An autopsy study. Annals of thoracic medicine, 12(3), 191–198. https://doi.org/10.4103/atm.ATM_276_16

[4] Kenfield, S. A., Wei, E. K., Stampfer, M. J., Rosner, B. A., & Colditz, G. A. (2008). Comparison of aspects of smoking among the four histological types of lung cancer. Tobacco control, 17(3), 198–204. https://doi.org/10.1136/tc.2007.022582

[5] Relli, V., Trerotola, M., Guerra, E., & Alberti, S. (2019). Abandoning the Notion of Non-Small Cell Lung Cancer. Trends in molecular medicine, 25(7), 585–594. https://doi.org/10.1016/j.molmed.2019.04.012

[6] Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., & Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nature genetics, 45(10), 1113–1120. https://doi.org/10.1038/ng.2764

[7] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics, 25(1), 25–29. https://doi.org/10.1038/75556

[8] Groelz, D., Sobin, L., Branton, P., Compton, C., Wyrich, R., & Rainen, L. (2013). Non-formalin fixative versus formalin-fixed tissue: a comparison of histology and RNA quality. Experimental and molecular pathology, 94(1), 188–194. https://doi.org/10.1016/j.yexmp.2012.07.002

[9] Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology, 15(12), 550. https://doi.org/10.1186/s13059-014-0550-8

[10] Sato, T., Soejima, K., Arai, E., Hamamoto, J., Yasuda, H., Arai, D., Ishioka, K., Ohgino, K., Naoki, K., Kohno, T., Tsuta, K., Watanabe, S., Kanai, Y., & Betsuyaku, T. (2015). Prognostic implication of PTPRH hypomethylation in non-small cell lung cancer. Oncology reports, 34(3), 1137–1145. https://doi.org/10.3892/or.2015.4082

[11] Zhong, X., Guan, X., Liu, W., & Zhang, L. (2014). Aberrant expression of NEK2 and its clinical significance in non-small cell lung cancer. Oncology letters, 8(4), 1470–1476.

https://doi.org/10.3892/ol.2014.2396

[12] Yu, G., Wang, L. G., Han, Y., & He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. Omics : a journal of integrative biology, 16(5), 284–287. https://doi.org/10.1089/omi.2011.0118

[13] Chen, M., Liu, X., Du, J., Wang, X. J., & Xia, L. (2017). Differentiated regulation of immune-response related genes between LUAD and LUSC subtypes of lung cancers. Oncotarget, 8(1), 133–144. https://doi.org/10.18632/oncotarget.13346

[14] Travis, W. D., Brambilla, E., Nicholson, A. G., Yatabe, Y., Austin, J., Beasley, M. B., Chirieac, L. R., Dacic, S., Duhig, E., Flieder, D. B., Geisinger, K., Hirsch, F. R., Ishikawa, Y., Kerr, K. M., Noguchi, M., Pelosi, G., Powell, C. A., Tsao, M. S., Wistuba, I., & WHO Panel (2015). The 2015 World Health Organization Classification of Lung Tumors: Impact of Genetic, Clinical and Radiologic Advances Since the 2004 Classification. Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer, 10(9), 1243–1260. https://doi.org/10.1097/JTO.0000000000000630

[15] Park, H. J., Cha, Y. J., Kim, S. H., Kim, A., Kim, E. Y., & Chang, Y. S. (2017). Keratinization of Lung Squamous Cell Carcinoma Is Associated with Poor Clinical Outcome. Tuberculosis and respiratory diseases, 80(2), 179–186. https://doi.org/10.4046/trd.2017.80.2.179

[16] Li, Q., Ran, P., Zhang, X., Guo, X., Yuan, Y., Dong, T., Zhu, B., Zheng, S., & Xiao, C. (2018). Downregulation of N-Acetylglucosaminyltransferase GCNT3 by miR-302b-3p Decreases Non-Small Cell Lung Cancer (NSCLC) Cell Proliferation, Migration and Invasion. Cellular physiology and biochemistry : international journal of experimental cellular physiology, biochemistry, and pharmacology, 50(3), 987–1004. https://doi.org/10.1159/000494482

[17] He, Y., Liu, R., Yang, M., Bi, W., Zhou, L., Zhang, S., Jin, J., Liang, X., & Zhang, P. (2021). Identification of VWF as a Novel Biomarker in Lung Adenocarcinoma by Comprehensive Analysis. Frontiers in oncology, 11, 639600. https://doi.org/10.3389/fonc.2021.639600

[18] Hu, M., Ye, L., Ruge, F., Zhi, X., Zhang, L., & Jiang, W. G. (2012). The clinical significance of Psoriasin for non-small cell lung cancer patients and its biological impact on lung cancer cell functions. BMC cancer, 12, 588. https://doi.org/10.1186/1471-2407-12-588

[19] Che, D., Wang, M., Sun, J., Li, B., Xu, T., Lu, Y., Pan, H., Lu, Z., & Gu, X. (2021). KRT6A Promotes Lung Cancer Cell Growth and Invasion Through MYC-Regulated Pentose Phosphate Pathway. Frontiers in cell and developmental biology, 9, 694071. https://doi.org/10.3389/fcell.2021.694071

[20] Schabath, M. B., & Cote, M. L. (2019). Cancer Progress and Priorities: Lung Cancer. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 28(10), 1563–1579. https://doi.org/10.1158/1055-9965.EPI-19-0221

[21] Broderick S. R. (2020). Adjuvant and Neoadjuvant Immunotherapy in Non-small Cell Lung Cancer. Thoracic surgery clinics, 30(2), 215–220. https://doi.org/10.1016/j.thorsurg.2020.01.001

[22] Steven, A., Fisher, S. A., & Robinson, B. W. (2016). Immunotherapy for lung cancer. Respirology (Carlton, Vic.), 21(5), 821–833. https://doi.org/10.1111/resp.12789

[23] Garon, E. B., Hellmann, M. D., Rizvi, N. A., Carcereny, E., Leighl, N. B., Ahn, M. J., Eder, J. P., Balmanoukian, A. S., Aggarwal, C., Horn, L., Patnaik, A., Gubens, M., Ramalingam, S. S., Felip, E., Goldman, J. W., Scalzo, C., Jensen, E., Kush, D. A., & Hui, R. (2019). Five-Year Overall Survival for Patients With Advanced Non–Small-Cell Lung Cancer Treated With Pembrolizumab: Results From the Phase I KEYNOTE-001 Study. Journal of clinical oncology : official journal of the American Society of Clinical Oncology, 37(28), 2518–2527. https://doi.org/10.1200/JCO.19.00934