Evaluation of the screening and diagnosis model for the lung adenocarcinoma protein function module based on WGCNA and machine learning

Jiankun Liang¹,a, Fei Zhai¹, Junting Min², Rongwu Xiang¹, Shen Xiao¹, Luhua Liang¹*

¹School of Medical Equipment, Shenyang Pharmaceutical University, Benxi, Liaoning, 117004, China
²Shenyang Pharmaceutical University Wuya College of Innovation, Shenyang, Liaoning, 110016, China
aLiangjiankun@syphu.edu.cn
*Corresponding author: liangluhua1988@163.com

Abstract: This paper aims to identify lung adenocarcinoma biomarkers through WGCNA and machine learning and construct a clinical diagnosis model for lung adenocarcinoma. We used the lung cancer protein expression data from the CPTAC database to conduct differentiation analysis, built a WGCNA network of lung adenocarcinoma samples and a WGCNA network of lung tumor samples and normal samples, and assessed the overlapped module of these two networks using machine learning. GO and KEGG abundance analysis was conducted to find proteins related to lung tumor, a correlation network for proteins in the overlapped module was created to mine the target protein biomarkers, and machine learning was used to create and analyze a screening model. There were 2317 differentially expressed proteins obtained from 213 lung tumor samples from the CPTAC database; the lung tumor network and the lung para-tumor tissue joint network had two overlapping modules; through PPI network optimization, we mined 11 protein biomarkers related to lung adenocarcinoma. Through verification based on data outside TCGA, we found that the lung adenocarcinoma diagnosis model constructed based on the 11 protein biomarkers had high accuracy and stability, showing clinical and biological significance.

1. Introduction
Lung cancer has risen as one of the most threatening diseases that claim human lives[1], accounting for the second largest cause of deaths among males and gaining increasing attention among female[2]. Lung cancer has two types: non-small cell lung carcinoma (NSCLC) and small cell lung carcinoma (SCLC). 80% patients with lung cancer suffer from the former type, and most are already in the middle or late stages when diagnosed, with a low five-year survival rate[3]. NSCLC can be divided into esophageal squamous cell carcinoma (ESCC), lung adenocarcinoma (LUAD) and large-cell undifferentiated carcinoma (UCUC). LUAD accounts for 50% of primary NSCLC, basically deriving from bronchial epithelial cells and growing along the alveolar wall[4]. The tumor is surrounded by rich blood vessels, strongly invasive and metastatic. The early stage is the optimal timing for treatment of LUAD; however, symptoms of LUAD are hard to observe in the early stage. Thus, finding biomarkers for early-stage LUAD and establishing a reliable LUAD diagnosis model is of great
importance.

In these years, as more proteomic datasets are built and biomarker screening algorithms improve, it has made much headway to use biomarkers as informative indicators for biostructures and functions to detect complicated diseases. It can effectively improve the diagnosis efficiency to build noninvasive diagnosis models for complicated diseases using biomarkers through mathematical methods. Studies have shown that serum tumor biomarkers like ProGRP, NSE and TPS has practical values in recognition, diagnosis, prognosis evaluation, effect monitoring and recurrence prediction of NSCLC. Besides, progress in liquid biopsy technique and clinical research advancements in NSCLC gene mutation targets and targeted drugs like Rova-T have provided new directions for exploration of early-stage clinical diagnosis and treatment of NSCLC[5].

Clinical Proteomic Tumor Analysis Consortium (CPTAC) has established a proteomic database based on analysis of protein abundance in tissue samples by mass spectrometry. This study includes mainly the following works: the protein expression data of LUAD samples and adjacent normal tissues samples in CPPTAC database were analyzed, the WGCNA (weighted gene co-expression network analysis) and machine learning methods were used to construct a protein model related to the prognosis of LUAD, biomarkers for early diagnosis of LUAD were selected and used as clinical indicators, using these biomarkers to establish the clinical diagnosis model of LUAD by machine learning method. It provides the basis for early diagnosis and targeted treatment of LUAD.

2. Materials and Methods

2.1. Data sources and pre-processing

In the present study, 111 LUAD samples and 102 normal lung tissue samples and clinical characteristics data of patients were downloaded from the CPTAC database (https://cptac-data-portal.georgetown.edu). The collected data had few missing values, and the missing values were complemented by the impute package by R language and the complemented data were normalized by the limma package. According to the number of samples in the LUAD dataset and the normal tissue dataset, R language was used to group the expression data.

2.2 Screening protein maker modules

With |logFC|>0.5 and p<0.05 as the standard, significantly differential proteins in the LUAD dataset and the normal tissue dataset were screened out, and the WGCNA method was used to perform clustering analysis on the differentially expressed proteins. A tumor network, i.e., a weighted gene co-expression network for the LUAD samples (Network A), was constructed to compute the correlation between the tumor network module and clinical characteristics, the Cox proportional-hazards model was used to calculate the correlation between the tumor network module feature genes and the survival risks, and corresponding modules in the tumor network related to survival were identified. A tumor - para-carcinoma joint weighted gene co-expression network (Network B) was established, the correlation between Network B and the type of samples was analyzed, and the overlapping modules between the modules with large correlations and the selected modules in Network A were identified.

2.3. Evaluating the overlapping modules by machine learning

The protein data obtained from the CPTAC database were randomly divided into two groups: one as the training set, and the other the test set. The training set was used to train the models built based on support vector machine (SVM), random forest, naïve Bayesian machine learning algorithms to evaluate the results of overlapping modules, and the test set was used to assess the models’ classification performance.

2.4 Screening protein markers

The clusterProfiler package was used to perform GO analysis and KEGG pathway enrichment analysis.
on the proteins in the overlapping modules that had good machine-learning evaluation results; and with \( p<0.05 \) and \( \text{FDR}<0.01 \) as the standard, proteins related to formation and development of tumors in the bioprocess and signaling pathways were selected and input into the STRING database for PPI analysis; and with the PPI network centrality \( \geq 5 \) as the standard, the LUAD biomarkers were selected.

2.5 Establishing the biomarker-based diagnosis model
The LUAD sample data from The Cancer Genome Atlas (TCGA) database were used to perform external validation of protein biomarkers. The samples were randomly divided into the training set and the test set. The training set was used to train diagnosis models created based on SVM, random forest and naïve Bayesian machine learning algorithms, and the test set was used to evaluate the capacity of the biomarkers in identifying LUAD.

3. Result

3.1 Screening of differentially expressed proteins
With \(|\log\text{FC}|>0.5\) and \( p<0.05 \) as the standard, 213 tumor samples collected from the CPTAC database were analyzed, and 2317 differentially expressed proteins were obtained, 1122 of which were up-regulated and 1195 were down-regulated (Figure 1).

3.2 Screening of protein maker modules

3.2.1 Construction of Network A — the weighted gene co-expression network of tumor samples.
A weighted gene co-expression network of tumor samples was built based on the WGCNA package in R, and the pickSoft-Threshold function was used to calculate the scale-free degree of the network (Figure 2). Generally, the minimum soft threshold obtained when the scale-free model fitting coefficient \( R^2>0.85 \) and there exists connectivity is taken as the optimal soft threshold[6]. In this study, \( \beta=4 \) was taken as the optimal soft threshold, based on which the network was divided into 11 modules, as shown in Figure 3, where the gray module represents proteins that could not be classified.

Figure 1. Volcano plot of differentially expressed proteins
Figure 2. Screening of soft thresholds of tumor samples

Cluster Dendrogram

Figure 3. Division of the original modules of the tumor samples and merge of similar samples
3.2.2 Correlation between the weighted gene co-expression network of tumor samples and non-time-related clinical data.

The Pearson correlation coefficient between the non-time-related clinical data of the 213 tumor samples (gender, age, height, weight, BMI, tumor size, tumor stage) and feature proteins in each module was calculated, and a larger absolute value of the Pearson coefficient would indicate closer connection between the module and the protein[6]. The results showed that the absolute values of the Pearson coefficient between all the modules and clinical traits were less than 0.7, which indicated that the relationship between non-time-related clinical characteristics and protein modules was not close. It means the correlation was not significant.

3.2.3 Survival analysis of weighted gene co-expression network modules of tumor samples.

With the survival time of patients with clinical symptoms and the survival state as the survival function, and with the 11 protein modules obtained in Section 2.2.1 as the variable, multi-factor Cox regression analysis was performed. Table 1 shows the meaningful results, where the green-yellow and purple modules were correlated to survival: the p value of the green-yellow module <0.05 and that of the purple module <0.15 (the p value was large because the multiple factors may work complementarily in regression).

| Module | coef | exp(coef) | se(coef) | z    | p(|z|) |
|--------|------|-----------|----------|------|-------|
| greenyellow | 5.5872 | 266.9770 | 2.8462 | 1.963 | 0.0496 |
| purple           | -6.8326 | 0.0011 | 4.3133 | -1.584 | 0.1132 |

3.2.4 Construction of Network B — the tumor-para-carcinoma joint weighted gene co-expression network.

The joint weighted gene co-expression network for tumor samples and the adjacent normal tissue samples (Network B) was constructed. Specifically, first, hierarchical clustering was performed on tumor samples and the normal tissue samples to calculate the fitting coefficient of the two sets of samples under different soft thresholds[6]. In this study, β=4 was taken as the joint soft threshold. Correlations between the feature proteins and the type of samples in different modules were calculated, as shown in Figure 3. As shown in the figure, the correlation coefficients of all modules were above 0.7, which means that the modules divided in Network B had good biological value.

```
Module-type relationships

MEturquoise  -0.95
MEpink       -0.94
MEMagenta    -0.92
MEyellow     -0.89
MEbrown      -0.86
MEpurple     -0.86
MEgreen      -0.70
MEblack      0.85
MERed        0.89
MEblue       0.92
MEgrey       0.97
```

Figure 4. Relationship between modules of Network B and the type of tissues
3.2.5 Overlapping of protein modules between Network A and Network B

Network A and Network B were jointly analyzed. The Fisher’s exact test was performed to calculate the p value between modules in Network A and Network B to evaluate the significance of overlapping[6], as shown in Figure 5, where a darker color indicates a higher degree of overlapping. As analyses above show, the green-yellow and purple modules in Network A is correlated to survival, and as shown in Figure 5, the green-yellow module in Network A and the brown module in Network B had overlapping areas, and the purple module in Network A and the pink module in Network B were highly overlapped. The proteins in the former overlapping was marked as Overlapping Module 1 and those in the latter as Overlapping Module 2. The proteins in these two overlapping modules were analyzed further.

![correspondence of Tumor set-specific and Tumor-Normal consensus modules](image)

Figure 5. Overlapping modules between Network A and Network B

3.3 Evaluating overlapping modules by machine learning

The 213 protein samples collected from the CPTAC database were divided into two sets randomly: one as the training set and the other as the test set. The training set was used to train models built by the SVM, random forest and naïve Bayesian machine learning algorithms to evaluate the results of overlapping modules, and the test set was used to evaluate the classification performance of the models [7]. Figure 6 shows the ROC curve for the internal verification result of the overlapping modules: the two overlapping modules had good screening effect in all the three models, with an AUC both larger than 0.9 (Table 2).

Therefore, the 73 overlapping proteins in these two modules could be taken as biomarkers in subsequent screening.
Figure 6. Machine learning ROC curves of two overlapping modules

Table 2. AUC of the two modules under the three machine learning methods

| Module                  | SVM-AUC | Random forest-AUC | Naïve Bayesian-AUC |
|-------------------------|---------|-------------------|--------------------|
| Overlapping Module 1    | 0.953   | 0.954             | 0.908              |
| Overlapping Module 2    | 0.981   | 0.953             | 0.996              |

3.4. Initial screening of protein markers

The cluster-Profiler package was used to perform GO and KEGG abundance analysis on the 73 overlapping proteins in the two overlapping modules. The GO abundance results were screened based on the standards of FDR<0.01 and p<0.05, and the KEGG abundance pathway was screened based on the standard of p<0.05 (part of the results is shown in Figures 3 and 4). GO entries and pathways related to tumor formation and development include the following: biological processes related to reactive oxygen reaction and arginine metabolism as well as tumor necrosis factor signal pathways in Overlapping Module 1, biological processes related to vascular formation and repair as well as NSCLC pathways in Overlapping Module 2. From these GO entries and pathways, 19 candidate biomarkers (protein) that were possibly related to LUAD formation and development were identified, which were CDH5, MMRN2, PECAM1, COL6A6, COL4A1, CLEC14A, LAMA5, TEK, COL4A3, COL4A2, BMPER, ARG1, RHOJ, TIE1, ANGPT2, MMP9, MET, FOS, PADI4.

Table 3. Part of the results of GO and KEGG abundance analysis of proteins in Overlapping Module 1

| ID          | Representative proteins | Description                     | p value  |
|-------------|-------------------------|---------------------------------|----------|
| GO:0006959  | BPI/BST1/FCN1/IGHM      | humoral immune response         | 4.78E-7  |
| GO:0001818  | MMP8/PGLYRP1/SRGN       | Negative regulation of cytokines| 5.13E-6  |
| GO:0034614  | ARG1/FOS/MET/MMP9       | Response of cells to active oxygen | 1.37E-5 |
Table 4. Part of the results of GO and KEGG abundance analysis of proteins in Overlapping Module 2

| ID          | Representative proteins | Description                               | p value   |
|-------------|-------------------------|-------------------------------------------|-----------|
| GO:0006527  | ARG1/PADI4              | Arginine decomposition                    | 5.12E-5   |
| GO:0006909  | FCN1/IGHM/MET/PTX3      | Phagocytosis                              | 2.91E-4   |
| hsa04657    | FOS/MMP9                | IL-17 signal pathway                      | 1.97E-3   |
| hsa04668    | FOS/MMP9                | Tumor necrosis factor signal pathway      | 2.78E-3   |
| hsa04380    | FCGR3B/FOS              | Osteoclast differentiation                | 3.62E-3   |

3.5 Screening of closely connected proteins using PPI network

The 73 proteins in the two overlapping modules were input to the STRING database to create the PPI network, and proteins with a degree ≥5 were selected, and the intersection set they had with the previously obtained 19 candidate biomarker proteins was obtained. At last, 11 biomarkers were obtained: MMP9, CDH5, PECAM1, COL4A1, LAMA5, TEK, MET, COL4A3, ARG1, TIE1, ANGPT2.

3.6 Evaluation of the LUAD screening model

The LUAD gene expression profiles were downloaded from The Cancer Genome Atlas (TCGA) database to perform external verification of the LUAD biomarker screening model. There were 594 samples in total, including 58 normal samples and 536 LUAD samples. All samples were randomly divided into a training set and a test set. The training set was used to train the LUAD diagnosis model created based on the 11 LUAD protein biomarkers screened out by SVM, random forest and naïve Bayesian machine learning algorithms. The test set was used to assess the model. Table 5 shows the AUC of the models built by the three methods (all above 0.9), and Figure 7 shows their ROC curves. External data-based verification shows that the LUAD diagnosis model built based on the 11 protein biomarkers had high accuracy and stability.

Table 5. AUC values of LUAD diagnosis models built based on biomarkers obtained based on three machine learning methods

| Constructing proteins | SVM-AUC | Random forest -AUC | Naïve Bayesian -AUC |
|-----------------------|---------|--------------------|---------------------|
| 11 biomarkers         | 0.950   | 0.932              | 0.984               |

Figure 7. Evaluation of LUAD screening models built based on the 11 biomarkers
4. Conclusions and Discussions

Based on the CPTAC database, this paper mined 11 LUAD screening and diagnosis protein biomarkers using the WGCNA and machine learning methods: MMP9, CDH5, PECAM1, COL4A1, COL4A3, LAMA5, TEK, TIE1, MET, ARG1, ANGPT2.

MMP9 is a matrix metalloproteinase, which can degrade IV collagen and gelatin to damage complete basilar membranes, so that the tumor cells can invade adjacent tissues and migrate to neighboring blood vessels. This is consistent with the strong invasion and metastasis of LUAD cells [8-9].

CDH5 (vascular endothelial cadherin) works to maintain stability of vascular endothelial cells. In normal conditions, it only does not express in normally expressed vascular endothelial cells, but has upregulated expression in tumor cells. Studies have revealed that CDH5 antibodies can inhibit tumor angiogenesis in vitro and thereby suppress growth and metastasis of tumor. Therefore, it can be taken as a target for detection and treatment of some tumors [10].

PECAM1 (platelet endothelial adhesion molecule-1) is correlated to bioprocesses like the survival and proliferation of cells, metastasis of tumor cells, and cell cycle. A study has revealed that its upregulated expression is correlated to metastasis and degradation of tumor, and has been used as an indicator for development of tumors [11].

COL4A1 and COL4A3 are both Type IV collagen. They are important components of the basilar membrane, mainly involved in tumor invasion and metastasis [12-13].

LAMA5 belongs to a laminin protein gene. It is also an important component of the basilar membrane and involved in tumor metastasis. Changes in the laminin proteins can affect the closeness between cells and thereby make it easier for spread of tumor cells [14].

TEK and TIE1 are two types of tyrosine kinase receptors, involved in angiogenesis and growth and metastasis of tumor cells. Studies found that inhibiting pathways related to TEK can suppress tumor growth and metastasis. Studies on TIE1 are few, and it is supposed that it may have similar functions to TEK [15-16].

MET (hepatocyte growth factor receptor) is correlated to bioprocesses including cell proliferation, tumor invasion and angiogenesis. Aberrant MET expression has been observed in many tumor tissues. Suppressing MET expression can, to some extent, inhibit tumor growth and metastasis [17].

Expression of ARG1 (arginase 1) has correlations to proliferation of some tumors (gastric carcinoma, NSCLC, prostatic cancer, etc.) [18-19].

ANGPT2 belongs to the angiopoietin family involved in angiogenesis. It helps growth of new blood vessels around the tumor to provide nutrients for tumor cells to grow and proliferate [20].

Most of these proteins are correlated to angiogenesis, tumor proliferation and metastasis, which is consistent with the feature that the LUAD tissues have rich vessels around and tumor cells are very likely to migrate.

Nonetheless, the present study has some shortcomings. First, the models built based on machine learning methods are black box models which cannot export specific equations for diagnosis of LUAD. Moreover, the screening of GO and KEGG abundance results is subjective. In the screening process, the selected are all known bioprocesses and pathways that are related to tumor growth and development, and it is likely to neglect some potential targets.

External verification by TCGA proves that the LUAD diagnosis model built based on the 11 selected protein biomarkers in this study has good accuracy and stability, thereby having clinical and biological importance.

References
[1] Yi Zhao, Meiqi Chen, Shuting Hao, et al. Mechanism of Anti-lung Cancer Effect of Xiao Chaihutang Based on Network Pharmacology[J]. Chinese Journal of Experimental Traditional Medical Formulae, 2020,26(09):208-214.
[2] Eugenia Bezzecchi, Mirko Ronzio, Valentina Semeghin, et al.NF-YA Overexpression in Lung Cancer: LUAD[J].Genes,2020,11(2).
[3] Zhao Chen, Christine M. Fillmore, Peter S. Hammerman, Carla F. Kim, Kwok-Kin Wong. Non-small-cell lung cancers: a heterogeneous set of diseases [J]. Nature Reviews Cancer, 2014, 14(8).

[4] Pengyi Hao, Kun You, Haozhe Feng, et al. Lung adenocarcinoma diagnosis in one stage [J]. Neurocomputing, 2020, 392.

[5] Fang Pan. Current Status of Diagnosis and Treatment for Small Cell Lung Cancer and Advances in the Treatment of Small Cell Lung Cancer with ROVA-T [D]. Chongqing Medical University, 2018.

[6] Zhihong Zhao. Exploring Function Gene Modules in Lung Adenocarcinoma with Weighted Gene Co-expression Network Analysis (WGCNA) [D]. Peking Union Medical College, 2017.

[7] Lihua Wang. Detecting Biomarkers of Alzheimer's Disease Based on Differential Gene Co-expression Network [D]. Shandong University, 2019.

[8] Yanhua Chang, Ying Cai, Bingqing Zou, et al. Expression and clinical significance of DEK and MMP-9 in gastric adenocarcinoma [J]. Chinese Journal of Clinical and Experimental Pathology, 2020(04):449-451.

[9] Ning Zhou, Xiaoshu Tan, Tingting Xu, et al. Expression of RECK and MMP9 in triple negative breast cancer and their relationship with clinical features and prognosis [J]. Oncology Progress, 2019, 17(16):1942-1945.

[10] Jie Ding. Preparation and functional study of human vascular endothelial cadherin CDH5 monoclonal antibody [D]. Suzhou University, 2016.

[11] Mingzhi Chen, Hong Yang, Caihong Wang, et al. Expression and significance of PECAM1 in cancer tissues of elderly patients with NSCLC [J]. Journal of Clinical Pulmonary Medicine, 2019, 24(04):614-617.

[12] CHEN Yao, WANG Wei, WANG Liang, et al. Expression and Clinical Prognosis of COL4A1 in Gastric Cancer Based on Bioinformatics Prediction [J]. Anti-tumor Pharmacy, 2019, 9(04):626-631+652.

[13] HU An-bang, LI Wei-qing, HONG Guo-dai, et al. Correlation between serum COL4A3 protein and clinicopathological features and prognosis in patients with non-small cell lung cancer (NSCLC) [J]. Fudan University Journal of Medical Sciences, 2019, 46(05):605-610.

[14] Diana Maltseva, Maria Raygorodskaya, Evgeny Knyazev, et al. Knockdown of the α5 laminin chain affects differentiation of colorectal cancer cells and their sensitivity to chemotherapy [J]. Biochimie, 2020, 174.

[15] MA Yi-Chao LU Yan JIA Yong-ping. Mechanism and significance of angiopoietin-Tie pathway in coronary heart disease [J]. Chinese Journal of General Practice, 2019, 17(08): 1379-1383.

[16] Huiting Sun, Huajun Zhou. Biological function of tyrosine kinase receptor tie [J]. Medical Recapitulate, 2010, 16(03):348-350.

[17] MingleiZhuo, Zhen Liang, Yuting Yi, et al. Analysis of MET kinase domain rearrangement in NSCLC [J]. Lung Cancer, 2020, 145.

[18] Na Chen. The Effects of Angiopoietin on the Proliferation, Invasion and Angiogenesis of Cervical Cancer [D]. Huazhong University of Science & Technology, 2016.

[19] Xuhui Wang, Tanghai Duan, Zhaoliang Su, et al. The expression of arginase 1 in peripheral blood of esophageal cancer patients and its clinical significance [J]. Immunological Journal, 2012, 28(11):972-975+980.

[20] Yan Luo, Yuquan Wei. The regulating roles of angiopoietins/TEK-2 in angiogenesis [J]. Chinese Journal of Medical Genetics, 2006(01):63-66.