Mathematical Biosciences and Engineering

MBE, 17(1): 737–746.
DOI: 10.3934/mbe.2020038
Received: 27 March 2019
Accepted: 15 September 2019
Published: 28 October 2019

http://www.aimspress.com/journal/MBE

Research article

Lung adenocarcinoma pathology stages related gene identification

Gaozhong Sun[1] and Tongwei Zhao[2,] *

[1] Department of Cardiothoracic Surgery, Zhejiang Provincial People's Hospital, People's Hospital of Hangzhou Medical College, Hangzhou 310014, China
[2] Department of Oncology, Zhejiang Provincial People's Hospital, People's Hospital of Hangzhou Medical College, Hangzhou 310014, China

* **Correspondence:** Email: twzhao1978@126.com, Tel: +86-18758032699.

Abstract: ***Objective*:** Lung cancer is a deadly disease with the highest 5-year survival rate. Lung adenocarcinoma is the main subtype of non-small cell lung cancer (NSCLC). Correct staging is critical as the basis of treatment. So, the identification of genes associated with pathologic stages of lung adenocarcinoma is helpful in understanding the pathological mechanism and designing targeted therapeutic drugs. ***Methods*:** Random forest was suitable for high-dimensional data to identify variables associated with the outcome. The variable importance-based selection method was used to rank the candidate genes associated with pathologic stages of lung adenocarcinoma. Univariate regression was used to analyze the relationship between gene expression and prognosis. The protein-protein interaction network was used to show the interactions among the identified genes. The identified genes functional enrichment analysis was performed by GSEA software. ***Results*:** Twelve genes significantly associated with pathologic stages of lung adenocarcinoma were identified by random forest analysis. Eight of these genes were found to play roles in survival of patients with lung adenocarcinoma. Among the 12 genes, 4 genes such as *CENPH*, *SRSF5*, *PITX2* and *NSG1* interacted with each other. And these genes were mainly enriched in p53 signaling pathway, cell cycle signaling pathway, JAK STAT signaling pathway and DNA replication signaling pathway. ***Conclusion*:** The identified genes may drive the changes among pathologic stages of lung adenocarcinoma and will be helpful in understanding the molecular changes underlying the pathologic stages.

Keywords: lung adenocarcinoma; gene; pathologic stages

1. Introduction

Lung cancer is a deadly disease with a 5-year survival rate of less than 15% and has two main subtypes which are known as non-small-cell lung carcinoma (NSCLC) and small cell lung carcinoma (SCLC) [1,2]. Lung adenocarcinoma is the main subtype of NSCLC. Previous studies have shown that lung adenocarcinoma patients have a shorter survival time than patients with another subtype of NSCLC [3].

Cancer staging is the process of determining the extent to which a cancer has developed by growing and spreading and provides a standardized framework to define the spread of a tumor [4]. Correct staging is critical as the basis of treatment (particularly the need for pre-operative therapy and/or for adjuvant treatment, the extent of surgery). Thus, incorrect staging would lead to improper treatment. Previous studies have suggested that lung cancer stages are associated with the prognosis of patients [5,6]. Patients in the early stages of lung cancer tend to live longer than those in later stages. Thus, it is important to identify genes associated with cancer staging which would be helpful in understanding the pathological mechanism and designing targeted therapeutic drugs.

Many gene selection methods are based on univariate rankings of gene relevance and arbitrary thresholds to select relevant genes. Most of these approaches can only be applied to two-class problems [7]. So, these methods do not suit to identify genes associated with pathologic stages of lung adenocarcinoma. Some methods can only distinguish genes with a linear correlation between expression levels and the outcome. However, the relationship between gene expression and outcome may be non-linear. Random forest is a machine learning method suited for classification in gene expression data [7,8]. It can be used when the number of observations is much smaller than the number of variables even when many of predictive variables are noise [7] and can be applied to problems containing more than two classes. And the returned variable importance is helpful in selecting the relevant genes which may be useful to predict the outcome.

In this study, we used the significance threshold of 2.52×10^{-5} (0.05 19879, Bonferroni correction) in random forest to identify genes that may lead to pathological staging of lung adenocarcinoma., and the molecular mechanism related to pathological staging was further explored by PPI analysis and GSEA enrichment analysis.

2. Materials and method

2.1. Gene expression dataset

The gene expression data of 572 patients and the corresponding clinical follow-up information were downloaded from the Cancer Genome Atlas (TCGA) databases. After removing those without pathologic stages information, totally 503 patients were included in this analysis. The RNA sequencing data of all patients' tumor tissues were standardized and then normalized within and among the left samples. We defined that a gene was abundantly expressed when its expression level was above 0 and appeared in at least 50% of the total samples.

2.2. Random forest

Random forest is a machine learning approach used for classification and regression developed

by Leo Breiman [9]. Based on a bootstrap sample of the data, each of the classification trees was split using a random subset of the candidate variables. To select the variables identified by random forest method, we used a computationally fast variable importance test method by conducting with standard procedure in R packages vita [10]. The previous study suggests that this method is a powerful approach suitable for high-dimensional data (e.g. gene expression data). Especially, when there are no predictor variables associated with the trait under a null model, the method vita is the most robust [10,11]. Unlike other permutation-based methods [12], vita approach only uses the existed data to estimate the null distribution of the variable importance [10]. In summary, the vita approach divides the whole data into two equally sized subsets, and two random forests are trained by using either of the two subsets. The variable importance is then calculated based on the other subset [10]. The variable importance showed the importance of each gene expression for classification into 4 stages. Finally, p values are calculated based on the distribution of variable importance [10].

2.3. Protein-protein interaction (PPI) network

To evaluate the interaction between identified gens, we extracted a small PPI network from the whole PPI network which was downloaded from cisPath database with the following link http://www.isb.pku.edu.cn/cispath/ [13]. The cisPath database contains PPI data from UniProt [14], PINA [15], iRefIndex [16] and STRING database [17]. The method SubNet was used to extract the small PPI network [18]. The subnetwork was then visualized the networks by Cytoscape [19].

2.4. Single gene functional enrichment analysis by GSEA

We performed Gene Set Enrichment Analysis (GSEA) with the software provided by the Massachusetts Institute of Technology [PMID：16199517]. GSEA is a computational method that determines whether a priori defined set of genes shows statistically significant, concordant differences between two biological states. In our study, the identified genes were analyzed by GSEA single gene functional enrichment analysis, and the enrichment results were shown by bubble diagram.

3. Results

Totally 503 patients were included in this analysis, after removing those without pathologic stages information. Details of clinical information were shown in Table 1. After removing those genes with low expression, totally 19,879 genes were contained in the analysis. So, the significant threshold of variable importance test was set to 2.52×10^{-5} (0.05/19879, Bonferroni correction). Based on this criterion, we identified 12 genes significantly associated with pathologic stages of lung adenocarcinoma. These genes with p value from the variable importance test were summarized in Table 2.

With the 12 genes, we constructed a pathologic stage classifier which could accurately predict the multi-stages of each patient with the area under the curve (AUC) of 0.899 using the random forest method. The receiver operating characteristic (ROC) curve for pathologic stage I was shown in Figure 1.

Table 1. Clinical information of 503 patients.

Characteristics	Summary
Lung adenocarcinoma patients	N = 503
Age	66.0±9.5 years
Gender	45.6% Male; 54.4% Female
Stage	
Stage Ⅰ	268 (53.3%)
Stage Ⅱ	121 (24.1%)
Stage Ⅲ	83 (16.5%)
Stage Ⅳ	26 (5.2%)

Table 2. Genes significantly associated with pathologic stages of lung adenocarcinoma.

Gene names	Variable importance	p value
ACSS1	2.12×10^{-4}	2.16×10^{-16}
LOC101928882	1.87×10^{-4}	2.19×10^{-16}
ZNF546	1.42×10^{-4}	2.21×10^{-16}
KCNC1	1.39×10^{-4}	2.21×10^{-16}
NSG1	1.38×10^{-4}	2.21×10^{-16}
PITX2	1.25×10^{-4}	2.22×10^{-16}
CENPH	1.25×10^{-4}	2.22×10^{-16}
BCAT2	1.21×10^{-4}	2.23×10^{-16}
MEI1	1.20×10^{-4}	2.23×10^{-16}
SRSF5	1.20×10^{-4}	2.23×10^{-16}
ZNF432	1.20×10^{-4}	2.23×10^{-16}
CNTN3	1.18×10^{-4}	2.23×10^{-16}

Then, we found that *ACSS1* ($p < 0.001$), *CNTN3* ($p = 0.009$), *BCAT2* ($p = 0.032$), *MEI1* ($p = 0.03$), *NSG1* ($p = 0.042$), *LOC101928882* ($p < 0.001$), *ZNF546* ($p = 0.037$) and *CENPH* ($p < 0.001$) were associated with prognosis by univariate regression analysis of 12 genes (Figure 2). Also, we found that *ACSS1*, *CENPH*, *CNTN3*, *KCNC1*, *MEI1*, *NSG1*, *SRSF5*, *ZNF432* and *ZNF546* were significantly differentially expressed in different pathological stages (Figure 3). The gene expression of *MEI1* and *ACSS1* decreased with the occurrence and development of lung adenocarcinoma, while the gene expression of *PITX2* and *CENPH* increased with the occurrence and development of lung adenocarcinoma.

We used SubNet to extract the small network from the whole PPI network. The result was displayed graphically in Figure 4. The genes obtained from PPI network were intersected with 12 genes screened by Random forest, and the four genes in the intersection were analyzed by GSEA single gene functional enrichment analysis, and the enrichment results were shown by bubble diagram as shown in Figure 5. Single gene functional enrichment analysis showed that *CENPH* and *SRSF5* were mainly enriched on P53 signaling pathway and cell cycle signaling pathway, *PITX2* was

mainly enriched on toll like receptor signaling pathway and JAK/STAT signaling pathway, and NSG1 was mainly enriched on JAK/STAT signaling pathway，DNA replication and cell cycle signaling pathway.

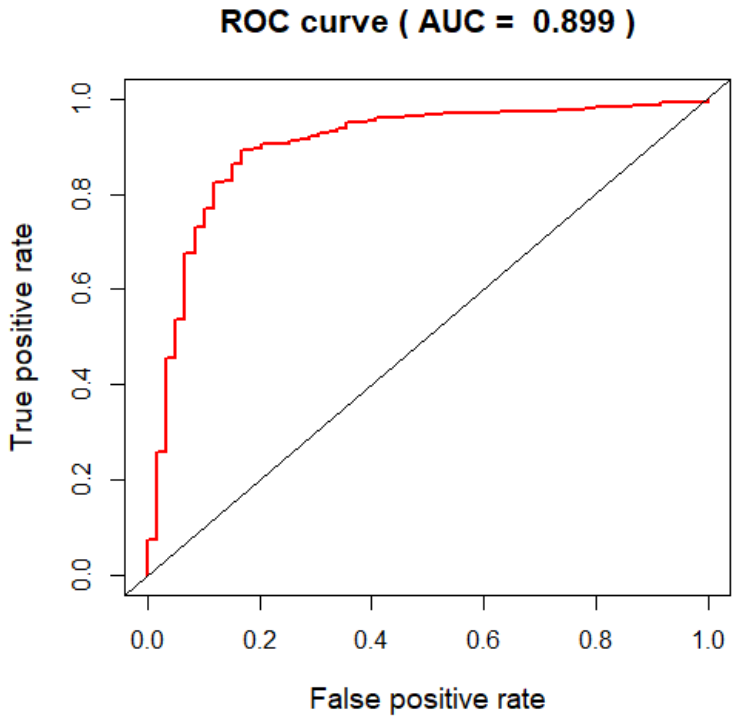

Figure 1. The ROC curve for pathologic stage Ⅰ.

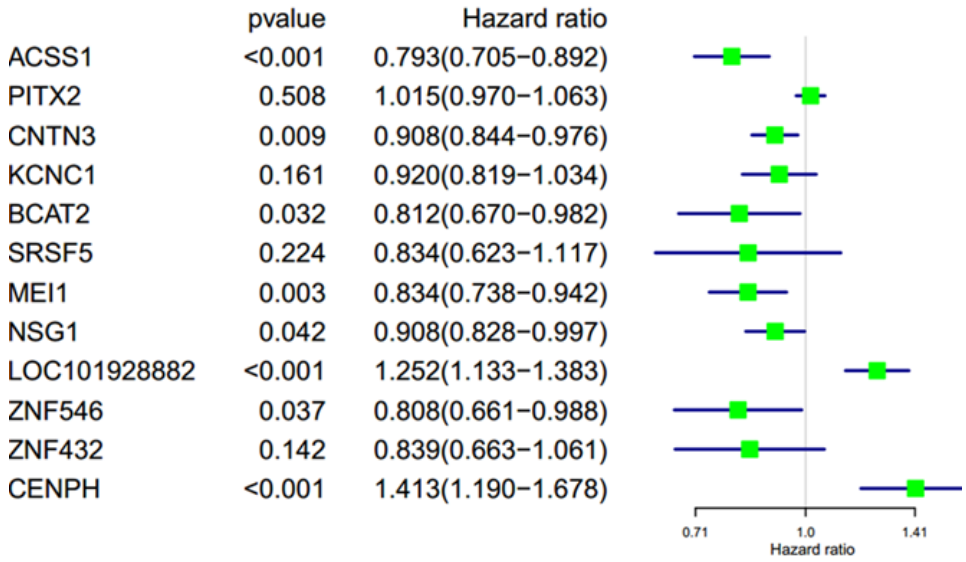

Figure 2. Regression analysis of univariate variables of 12 genes.

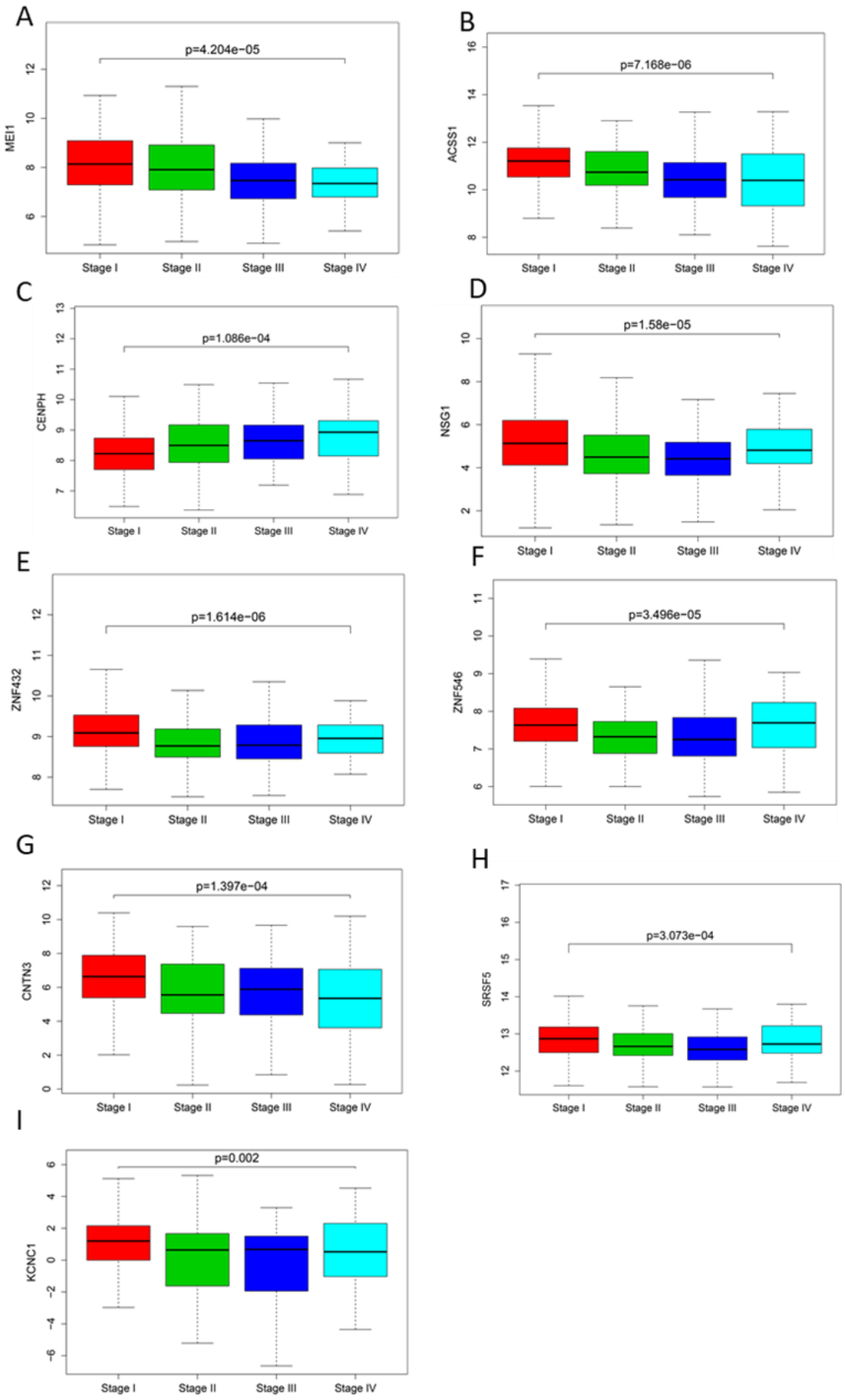

Figure 3. The 9 genes express differentially in different pathological stages. A: *MEI1*, B: *CSS1*, C: *CENPH*, D: *NSG1*, E: *ZNF432*, F: *ZNF546*, G: *CNTN3*, H: *SRSF5*, I: *KNC1*.

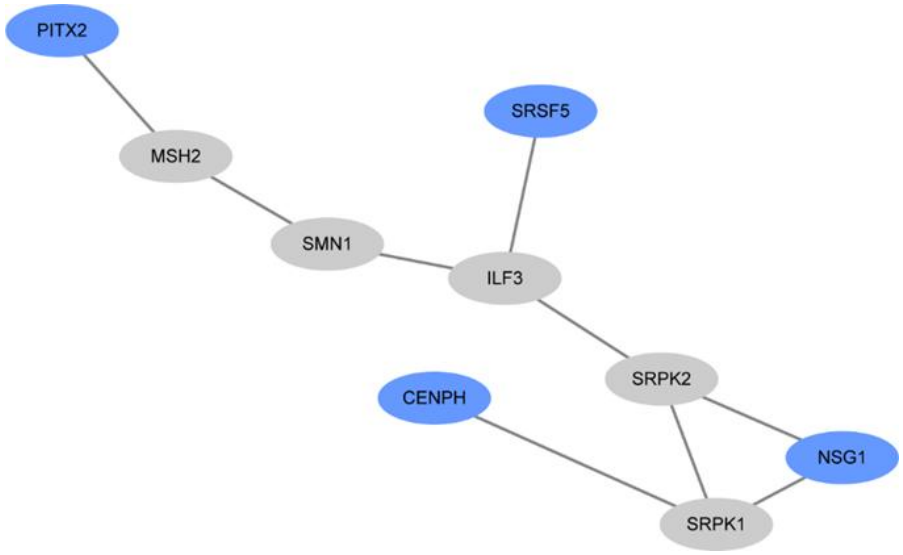

Figure 4. The extracted PPI network. Genes in blue were identified by random forest approach.

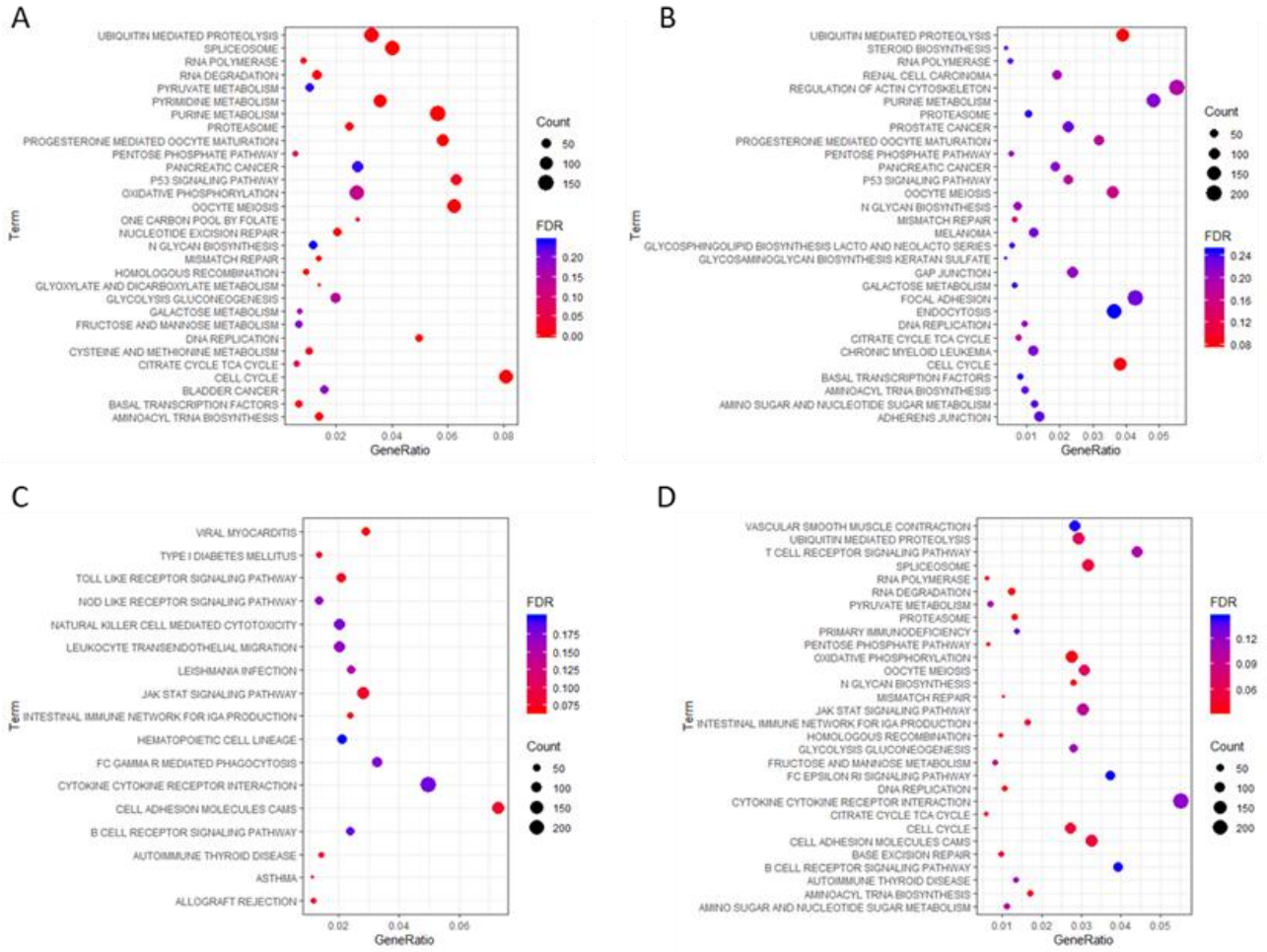

Figure 5. Gene function enrichment analysis. A: CENPH, B: SRSF5, C: PITX2, D: NSG1.

4. Discussion

In this study, we identified 12 genes significantly associated with pathologic stages of lung adenocarcinoma based on the random forest method. Many of them were found to play an import role in the tumor process, such as growth and metastasis.

Some of these genes were found to be differently expressed in tumor tissue compared with normal tissue. KCNC1 was found to be highly expressed in lung tumor tissue compared with normal tissues [20]. We found that the expression of KCNC1 was different in different cancer stages. NSG1 was identified as a biomarker for lung cancer early diagnosis [21]. We compared the expression of NSG1 in early stage (stage Ⅰ) and later stages and found that the expression of NSG1 was significantly higher in early stage than later stages (p value = 0.00023). This also suggested that NSG1 was associated with different cancer stages. On the other hand, single gene enrichment analysis of NSG1 was carried out by GSEA, and it was found that NSG1 was mainly enriched on JAK/STAT signaling pathway，DNA replication and cell cycle signaling pathway. It suggested that NSG1 may regulate the occurrence and development of lung adenocarcinoma through the above signaling pathways. In both head and neck region of squamous cell carcinoma and lung cancer patients, the methylation status of PITX2 was significantly associated with the survival status [22–24]. Since lung cancer stages are associated with the prognosis of patients [5,6], it is consistent with our results. CENPH protein, together with CENPA protein and CENPC protein, is a fundamental component of the active kinetochore complex [25,26]. The expression of CENPH was much higher in NSCLC cell lines and tumor tissues. Patients with a higher expression level of CENPH tended to have worse overall survival compared with patients with lower expression levels of CENPH [27], and our study found that the gene expression level of CENPH increased with the occurrence and development of lung cancer. The expression of SRSF5 was found to be highly expressed in tumor tissue. Knockdown of SRSF5 gene expression significantly reduced NSCLC cell proliferation [28]. In addition, using GSEA enrichment analysis, we found that SRSF5 may regulate the proliferation of lung adenocarcinoma cells through p53 signaling pathway and cell cycle signaling pathway.

Three of these genes were found to play important roles in the tumor process. With the help of acetate, glutamine supplementation is able to prevent loss of cell viability and support cell survival [29]. ACSS1, one type of acetyl-CoA synthetases, converted acetate to acetyl-CoA [30]. ACSS1 incorporates fatty acids into membranes to store energy and paly roles in metabolism [31]. However, short chain fatty acids cannot prevent the loss of cell viability from glucose deprivation [29]. In vivo studies also suggested that ACSS1 was a key factor in tumor growth in mice [29]. Our study suggested that ACSS1 was associated with different cancer stages, and the gene expression level of ACSS1 decreased gradually with the occurrence and development of lung adenocarcinoma. Depletion of BCAT2 was found to be able to impair the ability of lung tumor cell line to form a subcutaneous and orthotopic tumor in recipient mice [32].

5. Conclusion

In conclusion, we identified 12 genes associated with pathologic stages of lung adenocarcinoma by using random forest approach. Most of these genes were found to be important for cell proliferation or survival by previous studies. The left genes were warranted to find their significance.

These genes will be helpful in understanding molecular changes underlying the pathologic stages.

Conflict of interest

The authors declare no conflicts of interest.

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. D. S. Ettinger, W. Akerley, H. Borghaei, et al., Non-small cell lung cancer, version 2, 2013, *J. Natl. Compr. Canc. Netw.*, **11** (2013), 645–653, quiz 653.
2. R. Siegel, D. Naishadham and A. Jemal, Cancer statistics, 2013, *CA Cancer J. Clin.*, **63** (2013), 11–30.
3. D. E. Williams, P. C. Pairolero, C. S. Davis, et al., Survival of patients surgically treated for stage I lung cancer, *J. Thorac. Cardiovasc. Surg.*, **82** (1981), 70–76.
4. G. A. Woodard, K. D. Jones and D. M. Jablons, Lung Cancer Staging and Prognosis, *Cancer Treat. Res.*, **170** (2016), 47–75.
5. K. H. Yu, C. Zhang, G. J. Berry, et al., Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features, *Nat. Commun.*, **7** (2016), 12474.
6. F. C. Detterbeck, D. J. Boffa and L. T. Tanoue, The new lung cancer staging system, *Chest*, **136** (2009), 260–271.
7. R. Diaz-Uriarte and S. Alvarez de Andres, Gene selection and classification of microarray data using random forest, *BMC Bioinform.*, **7** (2006), 3.
8. V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, et al., Inferring regulatory networks from expression data using tree-based methods, *PLoS One*, **5** (2010).
9. L. Breiman, Random Forests, *Mach. Learn.*, **45** (2001).
10. S. Janitza, E. Celik and A. L. Boulesteix, A computationally fast variable importance test for random forests for high-dimensional data, *Adv. Data Anal. Classific.*, (2016).
11. F. Degenhardt, S. Seifert and S. Szymczak, Evaluation of variable selection methods for random forests and omics data sets, *Brief Bioinform.*, (2017).
12. A. Altmann, L. Tolosi, O. Sander, et al., Permutation importance: a corrected feature importance measure. *Bioinformatics*, **26** (2010), 1340–1347.
13. L. Wang, L. Yang, Z. Peng, et al., cisPath: an R/Bioconductor package for cloud users for visualization and management of functional protein interaction networks, *BMC Syst. Biol.*, **9** (2015), S1.
14. A. Noto, S. Raffa, C. De Vitis, et al., Stearoyl-CoA desaturase-1 is a key factor for lung cancer-initiating cells, *Cell Death Dis.*, **4** (2013), e947.
15. M. J. Cowley, K. S. Pinese, N. Kassahn, et al., PINA v2.0: Mining interactome modules, *Nucleic Acids Res.*, **1** (2012), 40.
16. S. Razick, I. M. Magklaras and I. M. Donaldson, iRefIndex: A consolidated protein interaction

database with provenance, *BMC Bioinform.*, **9** (2008), 405.
17. A. Franceschini, S. Szklarczyk, M. Frankild, et al., STRING v9.1: Protein-protein interaction networks, with increased coverage and integration, *Nucleic Acids Res.,* **11** (2013), 41.
18. C. Lemetre, Q. Zhang and Z. D. Zhang, SubNet: A Java application for subnetwork extraction, *Bioinformatics*, **29** (2013), 2509–2511.
19. P. Shannon, O. Markiel, N. S. Ozier, et al., Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.,* **13** (2003), 2498–2504.
20. Y. J. Kwon, S. J. Lee, J. S. Koh, et al., Genome-wide analysis of DNA methylation and the gene expression change in lung cancer, *J. Thora. Onco.*, **7** (2012), 20–33.
21. J. Pan, G. Song, D. Chen, et al., Identification of serological biomarkers for early diagnosis of lung cancer using a protein array-based approach, *Mol. Cell. Proteom.*, (2017).
22. E. E. Holmes, M. Jung, S. Meller, et al., Performance evaluation of kits for bisulfite-conversion of DNA from tissues, cell lines, FFPE tissues, aspirates, lavages, effusions, plasma, serum, and urine, *PLoS One*, **9** (2014), e93933.
23. V. Sailer, E. E. Holmes, H. Gevensleben, et al., PITX3 DNA methylation is an independent predictor of overall survival in patients with head and neck squamous cell carcinoma, *Clin. Epigenet.*, **9** (2017), 12.
24. V. Sailer, E. E. Holmes, H. Gevensleben, et al., PITX2 and PANCR DNA methylation predicts overall survival in patients with head and neck squamous cell carcinoma. *Oncotarget*, **7** (2016), 75827–75838.
25. T. Fukagawa, A. Mikami, V. Nishihashi, et al., CENP-H, a constitutive centromere component, is required for centromere targeting of CENP-C in vertebrate cells, *EMBO J.*, **20** (2001), 4603–4617.
26. N. Sugata, W. C. Li, T. J. Earnshaw, et al., Human CENP-H multimers colocalize with CENP-A and CENP-C at active centromere--kinetochore complexes, *Hum. Mol. Genet.*, **9** (2000), 2919–2926.
27. W. T. Liao, X. Wang, L. H. Xu, et al., Centromere protein H is a novel prognostic marker for human nonsmall cell lung cancer progression and overall patient survival, *Cancer*, **115** (2009), 1507–1517.
28. H. R. Kim, G. O. Lee, K. H. Choi, et al., SRSF5: A novel marker for small-cell lung cancer and pleural metastatic cancer, *Lung Cancer*, **99** (2016), 57–65.
29. A. J. Lakhter, J. Hamilton, R. L. Konger, et al., Glucose-independent acetate metabolism promotes melanoma cell survival and tumor growth, *J. Biol. Chem.*, **291** (2016), 21869–21879.
30. L. K. Boroughs and R. J. DeBerardinis, Metabolic pathways promoting cancer cell survival and growth, *Nat. Cell. Biol.*, **17** (2015), 351–359.
31. P. A. Watkins, D. Maiguel, Z. Jia, et al., Evidence for 26 distinct acyl-coenzyme A synthetase genes in the human genome, *J. Lipid. Res.*, **48** (2007), 2736–2750.
32. E. M. Kerr and C. P. Martins, Metabolic rewiring in mutant Kras lung cancer, *FEBS J.*, **285** (2018), 28–41.