Validation Study

Journal of
INTERNATIONAL
MEDICAL RESEARCH

Journal of International Medical Research
48(3) 1–12
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0300060519893837
journals.sagepub.com/home/imr
SAGE

Weighted gene expression profiles identify diagnostic and prognostic genes for lung adenocarcinoma and squamous cell carcinoma

Xing Wu[1],*, Linlin Wang[2],*, Fan Feng[3] and Suyan Tian[4]

Abstract

Objective: To construct a diagnostic signature to distinguish lung adenocarcinoma from lung squamous cell carcinoma and a prognostic signature to predict the risk of death for patients with nonsmall-cell lung cancer, with satisfactory predictive performances, good stabilities, small sizes and meaningful biological implications.

Methods: Pathway-based feature selection methods utilize pathway information as a priori to provide insightful clues on potential biomarkers from the biological perspective, and such incorporation may be realized by adding weights to test statistics or gene expression values. In this study, weighted gene expression profiles were generated using the GeneRank method and then the LASSO method was used to identify discriminative and prognostic genes.

Results: The five-gene diagnostic signature including keratin 5 (*KRT5*), mucin 1 (*MUC1*), triggering receptor expressed on myeloid cells 1 (*TREM1*), complement C3 (*C3*) and transmembrane serine protease 2 (*TMPRSS2*) achieved a predictive error of 12.8% and a Generalized Brier Score of 0.108, while the five-gene prognostic signature including alcohol dehydrogenase 1C (class I), gamma polypeptide (*ADH1C*), alpha-2-glycoprotein 1, zinc-binding (*AZGP1*), clusterin (*CLU*), cyclin dependent kinase 1 (*CDK1*) and paternally expressed 10 (*PEG10*) obtained a log-rank *P*-value of 0.03 and a C-index of 0.622 on the test set.

[1]Department of Teaching, The First Hospital of Jilin University, Changchun, Jilin Province, China
[2]Department of Ultrasound, China-Japan Union Hospital of Jilin University, Changchun, Jilin Province, China
[3]School of Mathematics, Jilin University, Changchun, Jilin Province, China
[4]Division of Clinical Research, The First Hospital of Jilin University, Changchun, Jilin Province, China

*These authors contributed equally to this work.

Corresponding author:
Suyan Tian, Division of Clinical Research, The First Hospital of Jilin University, 71 Xinmin Street, Changchun, Jilin Province 130021, China.
E-mail: windytian@hotmail.com

Conclusions: Besides good predictive capacity, model parsimony and stability, the identified diagnostic and prognostic genes were highly relevant to lung cancer. A large-sized prospective study to explore the utilization of these genes in a clinical setting is warranted.

Keywords

Lung adenocarcinoma, squamous cell carcinoma, GeneRank, prognosis, classification, pathway information

Date received: 8 August 2019; accepted: 19 November 2019

Introduction

Nonsmall-cell lung cancer (NSCLC) is any type of epithelial lung cancer (LC) other than small-cell lung carcinoma (SCLC) and it accounts for approximately 85% of all LC cases.[1] Compared with SCLC, patients with NSCLC are relatively less sensitive to chemotherapy.[2–4] In contrast to NSCLC, SCLC has a shorter doubling time, higher growth fraction and earlier development of metastases.[4,5] Furthermore, NSCLC can be classified into three major histological subtypes: lung adenocarcinoma (AC), lung squamous cell carcinoma (SCC) and large cell lung cancer (LCLC);[1] accounting for approximately 4/5 of all LC cases when combined together.[6] Since the choice of chemotherapy and targeted therapies depends on histological subtypes, the discrimination and separation of NSCLC subtypes are of essential importance in the clinical setting.[7] Likewise, the successful prediction of which patients with NSCLC have a high risk for recurrence and death is of primary importance with regard to the provision of more individualized and precise medical interventions.

A gene signature is a list of genes with a unique pattern of gene expression that results from an altered biological process and/or a medical condition.[8,9] According to the type of outcomes, a gene signature can be classified into either a diagnostic signature or a prognostic one. A diagnostic gene signature might provide valuable clues on biomarkers that distinguish the patients with a specific disease from the healthy controls; or different diseases with phenotypically similar medical conditions.[10] In contrast, a prognostic signature may offer insights into the course of a disease, the prediction of survival rates and the response to a specific treatment. [10,11] As far as NSCLC is concerned, many diagnostic signatures[12–15] and prognostic signatures[16–20] have been identified. For example, a recent study used 183 AC and 80 SCC patients as a training set and obtained a 42-gene signature to discriminate these two subtypes.[12] Furthermore, a 72-gene prognostic signature was shown to predict the risk of recurrence for early-stage NSCLC patients.[21]

Identification of a gene signature is usually accomplished with the aid of a feature selection process. Feature selection has the advantages of simplifying the final models, shortening the training time, alleviating the over-fitting problem and thus improving generalization and having a better biological interpretation. Different from the conventional feature selection methods that ignore biological pathway information or the underlying correlation or intrinsic grouping structure among genes, there exist many feature selection methods that incorporate such information to guide which genes should be selected. These methods branch out as a novel type of feature selection, dubbed as the pathway-based feature/gene selection.[22,23]

Previous studies have demonstrated that taking informative pathway knowledge into account, the pathway-based feature selection algorithms outperform the classic methods.[22,24–26] Consequently, such methods tend to replace the classic methods as the first choice of statistical methods in many real-world applications. So far, a majority of identified gene signatures for NSCLC were obtained by utilizing classic feature selection methods (e.g. a Cox model or a logistic model plus a LASSO penalty) that consider no pathway information.[27,28] To develop more pathway-based feature selection methods and then use them to construct better-performed gene signatures for NSCLC is highly desirable.

The GeneRank method[29] modified Google's PageRank algorithm[30] to specifically handle biological data. This method calculates a rank (i.e. GeneRank) for each gene and balances between the mean expression value or the fold change among phenotypes and the connectivity level of a gene within a network. GeneRank prioritizes a gene that is highly connected to other genes within the network. Previously, the GeneRank method was utilized to rearrange genes accordingly to their GeneRanks and restrict the search space (i.e. the number of genes under consideration) to those top-ranked genes. Then the corresponding *P*-values of Cox-filter models were adjusted according to a specific gene's correlation magnitudes with other genes involved in this search space, in order to eliminate redundant genes as much as possible.

The weighting methods are one major category of pathway-based feature selection methods,[13,22] but they have been underutilized since weight estimation is always subject to bias. In this current study, the GeneRank method[29] was applied to the expression values of genes to obtain weighted gene expression profiles and then the LASSO method[31] was used to select relevant genes, with the objective of identifying a diagnostic gene signature for subtype segmentation and a prognostic gene signature to predict overall survival rate. The resulting weighted gene expression profiles (i.e. the GeneRanks) bypassed the step of estimating pathway-based weights and thus provided an alternative strategy of weighting.

Notably, even though the GeneRank method was used in both this current study and our previous study,[32] the purposes of its implementation (to provide the rankings for genes versus to generate weighted expression profiles) and the feature selection methods used (the Cox-filter method versus the LASSO method) in both studies differ dramatically. Lastly, the objectives of these two studies were different. While our previous study focused on identifying subtype-specific prognostic genes, the current study aimed at a diagnostic gene signature to separate AC from SCC and a prognostic gene signature that works well for both subtypes.

Materials and methods

Experimental data

The data from three microarray experiments were used in this study. The accession numbers of the three microarray experiments in the Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) repository are GSE30219, GSE37745 and GSE50081. The RNA-Seq data of TCGA LUAD (for the AC subtype) and LUSC (for the SCC subtype) cohorts were used as an independent set to validate the performance of the resulting gene signatures. Since the three microarray datasets and the RNA-Seq dataset had been used by us previously,[32] the descriptions of them are not shown here. Since the same pre-processing procedures in that study were used, no description of the pre-processing procedures is given here. The readers are referred to our previous study for those details.[32]

The interaction/connection information of protein-coding genes was retrieved from the Human Protein Reference Database (HPRD),[33] Release 9 (http://www.hprd.org). There were 8023 unique genes commonly annotated by the R Hgu133plus2.db package, the RNA-seq data and the HPRD database, upon which the proposed procedure was applied.

Statistical analyses

GeneRank. The GeneRank method[29] calculates a rank for each gene that balances between its expression value and its importance within a gene-to-gene interaction network. Detailed descriptions of this method have been provided previously.[29,32]

Lasso. In order to be consistent with our previous study on the weighting methods,[13] the LASSO method was used to carry out the feature selection.[31] For two-class classification analysis, the corresponding log likelihood function and the final objective function were given in that previous article.[13] Here, the detailed description of the LASSO method for survival analysis was given, in which the following log partial likelihood function was used:

$$\log\Big(pl(\beta)\Big) = \sum_{i=1}^{n} \delta_i\Big(\beta^T X_i - \log\Big(\sum_{k \in r_i} \exp(\beta^T X_k)\Big)\Big)$$

Where δ_i is an event indicator, taking the value of 1 if the event happened, 0 if otherwise. If δ_i=1, t_i corresponds to the survival time (time free of the event) of subject *i*, otherwise t_i corresponds to the censoring time. Then, r_i indexes the risk set of patients for deaths at the moment of t_i. X_i=(X_{i1}, ..., X_{ip}) represent the expression measures of subject *i* for all genes under consideration.

Then the LASSO penalty term multiplied with a tuning parameter λ, i.e., $\lambda|\beta|_1$ is added to the negative log partial likelihood function to generate the final objective function, given as shown below:

$$-\log\Big(pl(\beta)\Big) + \lambda \sum_{j=1}^{p} |\beta_j|$$

Here, λ controls the sparseness of the final model, with a larger λ imposing heavy penalization on these β coefficients and a value of 0 corresponds to no penalty at all. The cyclic coordinate descent method implemented by the R glmnet package was used to optimize the final objective functions.[34]

Proposed procedure. The current study used the expression profiles of an individual directly to calculate the GeneRanks, which are viewed as patients' weighted gene expression profiles. Subsequently, the LASSO method was applied to select genes associated with subtype segmentation or with overall survival time upon the resulting weighted gene expression values, respectively. Based on the characteristics of the proposed procedure, it belongs to the weighting category described previously.[22] In that previous study,[22] a weighting method was defined to generate weights according to pathway information and then to combine those weights with either test statistics or expression values directly to accomplish the selection of relevant genes. Since the proposed procedure bypasses the step of estimating pathway-based weights, it provides an efficient alternative of weighting.

Performance statistics. For the classification, two metrics, the Generalized Brier Score (GBS) and the misclassification error rate, were used to evaluate the performance of a resulting gene signature. Detailed descriptions of these two metrics have been presented in a previous study.[22] In short, the closer to 0 these two statistics are the better a model performs.

Another two metrics were used to evaluate the performance of the resulting prognostic signatures. The log-rank *P*-value was the first metric. Briefly, the patients were stratified into two groups according to the mean of their risk scores: the low risk of death group and the high risk of death group. The survival curves of these two groups were plotted using the Kaplan–Meier method and compared by carrying out log-rank tests. The smaller a log-rank *P*-value was the more significantly these two survival curves differ. The second metric used was the censoring-adjusted C-statistic described previously.[35] For this metric, a value closer to 1 corresponds to a better performance.

In addition to the predictive performance of a gene signature, model stability and biological implication are also of crucial importance. If a specific statistical model identifies different gene lists on different data, its stability is low. If so, the application of resulting signatures in the clinical setting is impossible. The stability of a gene signature was evaluated using the bagging method.[36] Specifically, 100 bootstrapped replicates were generated by randomly sampling 4/5 of patients without replacements and then the percentages of the selected genes being inside these bagging signatures were calculated. The genes with low stability were discarded and the performance statistics were recalculated. Statistical analyses were performed using the R language (www.r-project.org) version 3.3.

Results

Using the weighted expression profiles generated by the GeneRank method, a 14-gene list was identified for the segmentation between AC and SCC. The estimated coefficients and their frequencies of being selected over the 100 bootstrapped datasets are given in Table 1, along with their biological relevance on the basis of the GeneCards database search.[37] The performance statistics of this 14-gene signature are listed in Table 2. Restricting the bagging frequencies of genes being selected to above 80%, the resulting five-gene list including keratin 5 (*KRT5*), mucin 1 (*MUC1*), triggering receptor expressed on myeloid cells 1 (*TREM1*), complement C3 (*C3*) and transmembrane serine protease 2 (*TMPRSS2*) obtained an error rate of 12.8% and a GBS of 0.108 on the test set.

Restricting further on the genes directly related to lung cancer, *KRT5*, *MUC1* and *TREM1* were kept. Interestingly, the

Table 1. The discriminative gene list to discriminate lung adenocarcinoma from lung squamous cell carcinoma.

Gene symbols	β	Percentage (%)	Biological relevance
ADH7	0.8913	53	I
C3	−1.1502	87	I
CALML3	0.4041	37	I
CHGA	0.0822	31	D
EGFR	−1.3647	73	D
GGH	1.0274	63	D
ISL1	2.4164	54	I
KRT5	9.6217	100	D
MSLN	–0.2808	51	D
MUC1	–3.2149	93	D
S100A7	0.4458	53	D
SFTPB	–0.2771	59	D
TMPRSS2	–2.3844	90	I
TREM1	–3.0412	99	D

β is the estimated coefficient for the specific gene (the magnitude of association with the outcome) using the LASSO model; percentage (%) is the frequency of being identified as a non-zero β over 100 replicates.
I, indirectly related to nonsmall-cell lung cancer according to the GeneCards database; D, directly related to non-small-cell lung cancer according to the GeneCards database.

Table 2. Performance statistics for both discriminative and prognostic gene signatures.

	Classification (AC versus SCC)		Prognosis (AC & SCC)	
	Error rate (%)	GBS	*P*-value (log-rank)	C-index
Training set (integrated microarray data)	6.49	0.056	8.05×10^{-9}	0.667
Test set (RNA-seq data)	14.4	0.109	0.252	0.577
Using bagging to eliminate the genes with low frequencies*				
Training set (integrated microarray data)	11.50	0.076	3.01×10^{-5}	0.630
Test set (RNA-seq data)	12.80	0.108	0.03	

*For the classification problem the genes with frequencies of > 80% were kept; and for the prognosis problem the genes with frequencies of > 50% were kept.

AC, lung adenocarcinoma; SCC, lung squamous cell carcinoma.; GBS, Generalized Brier Score.

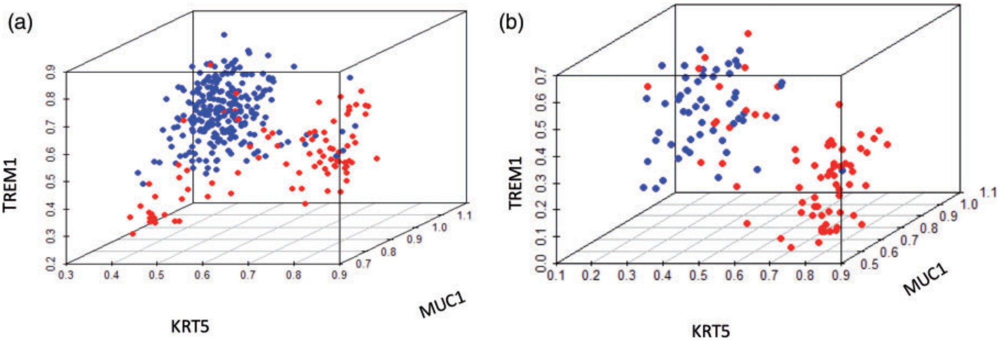

Figure 1. Scatterplots of the discriminative gene signature. (a) The training set (integrated microarray data). (b) The test set (RNA-Seq data). The three genes under consideration are keratin 5 (*KRT5*), mucin 1 (*MUC1*) and triggering receptor expressed on myeloid cells 1 (*TREM1*) that not only have high stability (the frequencies of being selected are > 80%) but also are directly related to nonsmall-cell lung cancer. From these two plots, the AC patients (blue dots) and the SCC patients (red dots) were observed to be well separated using these three genes. The weighted gene expression values for *KRT5* are given on the x-axis, for *MUC1* on the y-axis and for *TREM1* on the z-axis. AC, lung adenocarcinoma; SCC, lung squamous cell carcinoma. The colour version of this figure is available at: http://imr.sagepub.com.

bagging frequencies of these three genes were above 90% and their estimated association magnitudes (βs) were very large, while this three-gene list achieved a predictive error of 15.2% and a GBS of 0.101 on the test set. The scatterplots of the three-gene list are presented in Figure 1, upon which it was observed that AC and SCC patients were well separated into two clusters with only a small proportion of misclassifications for both the training set and the test set. In order to further validate the resulting signature, three sets of 1000 randomly-chosen gene lists were generated: one list included 14 genes; one included five genes with bagging frequencies > 80%; and the last list included three genes with bagging frequencies > 90%. Among 1000 randomly-chosen lists, none of them had better performances than the corresponding gene lists selected by the proposed procedure regarding error rates and GBS.

Using the weighted expression profiles generated by the GeneRank method,

Table 3. The prognostic gene list for nonsmall-cell lung cancer.

Gene symbols	β	Percentage (%)	Biological relevance
ADH1C	–0.3811	89	D
AZGP1	–0.7664	87	I
CD79A	–0.4586	46	I
CDK1	0.8026	65	D
CLU	−1.1324	66	D
COL10A1	–0.7247	47	I
GFRA3	–0.9063	49	I
GJB2	0.4625	44	D
MAOA	–0.3124	50	D
PEG10	0.5301	56	D
S100A7	0.0574	31	D
SCGB3A2	–0.089	45	I
SMAD9	–0.1602	44	D
TFF3	–0.013	35	I

β is the estimated coefficient for the specific gene (the magnitude of association with the outcome) using the LASSO model; percentage (%) is the frequency of being identified as a non-zero β over 100 replicates.
D, directly related to nonsmall-cell lung cancer according to the GeneCards database; I, indirectly related to nonsmall-cell lung cancer according to the GeneCards database.

another 14 genes were deemed to have prognostic values for NSCLC. The performance statistics of this prognostic gene signature are presented in Table 2. The estimated coefficients and their frequencies of being selected across 100 bootstrapped datasets are given in Table 3.

Using the weighted expression profiled by the GeneRank method both performance statistics of the resulting prognostic gene signature after eliminating the genes with low stability (i.e. a five-gene signature consisting of alcohol dehydrogenase 1C [class I], gamma polypeptide [*ADH1C*], alpha-2-glycoprotein 1, zinc-binding [*AZGP1*], clusterin [*CLU*], cyclin dependent kinase 1 [*CDK1*] and paternally expressed 10 [*PEG10*]) achieved satisfactory levels, namely, a C-index of 0.622 and a log-rank *P*-value of 0.03 on the test set. This result has two implications. First, the values of these two metrics are comparable with those on the training set, indicating no over-fitting occurs in this study and the learning model is well behaved. Secondly, a gene signature obtained from one platform such as microarray may be even generalized to another platform such as RNA-Seq.

In addition, a comparison was undertaken between the three-gene diagnostic signature and the gene lists given by two previous studies.[38,39] By fitting an extra support vector machine model using the identified genes of those two studies as covariates, the corresponding error rate and GBS on the test set were calculated, respectively. For the model using *KRT5*, RAR related orphan receptor C (*RORC*) and MAGE family member A4 (*MAGEA4*) as covariates, the corresponding error rate and GBS were 15.2% and 0.122 on the test set, respectively. In contrast, the corresponding values were 17.6% and 0.154 for the model using only *KRT5* as a covariate. Likewise, the 14-gene prognostic signature was compared with the prognostic gene sets given by two other previous studies.[40,41] The previous 13-gene list achieved a log-rank *P*-value of 0.09 and a C-index of 0.624 on the test set. On the other hand, the previous 15-gene list only obtained a log-rank *P*-value of 0.45 and a C-index of 0.514.

The Kaplan–Meier plots are presented in Figure 2. In order to compare with a random set of genes, two sets of 1000 randomly-chosen gene lists were generated (one included 14 genes and one included five genes with bagging frequencies > 50%). Among 1000 randomly-chosen 14-gene lists, none had more significant log-rank *P*-values and better C-indexes than the 14-gene list identified by the proposed procedure. For 1000 randomly-selected five-gene lists, 15 and 30 had more significant log-rank *P*-values and better C-indexes

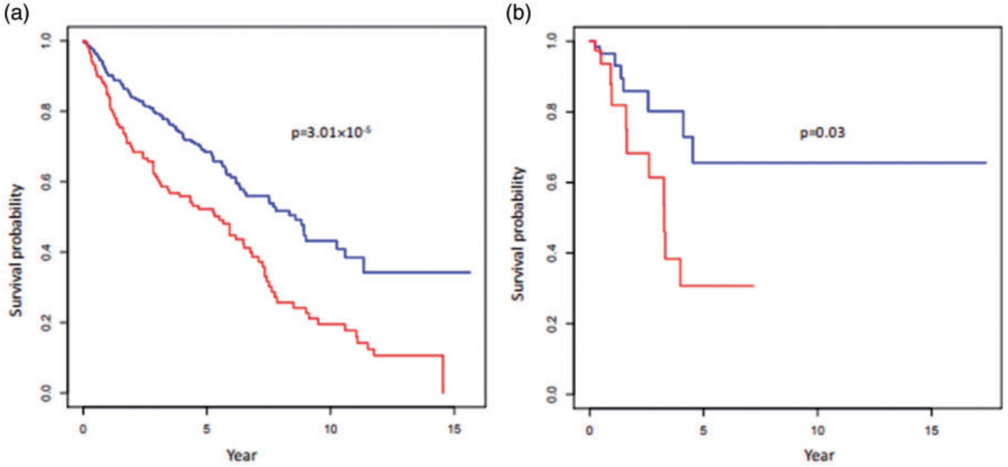

Figure 2. Kaplan–Meier plots for the five-gene prognostic signature. (a) The training set (the integrated microarray dataset). (b) The test set (the RNA-Seq dataset). Using the mean of risk scores as a cutoff, the patients were divided into two groups, i.e. the high-risk group (red solid line) and the low-risk group (blue solid line) and then a log-rank test was conducted to test if the survival curves of these two groups differed. *P*-value is the corresponding log-rank *P*-value. The colour version of this figure is available at: http://imr.sagepub.com.

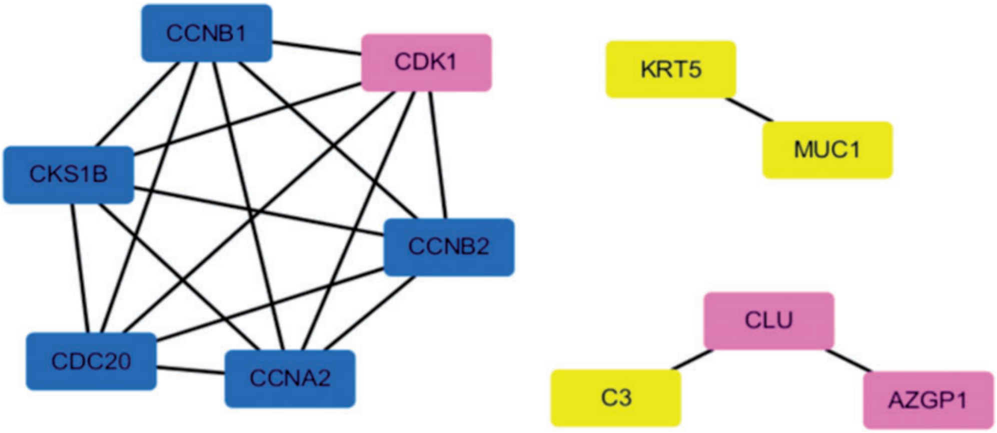

Figure 3. Interaction network on the basis of five prognostic genes and five diagnostic genes. In this graph, the isolated genes were excluded. The identified diagnostic genes (i.e. keratin 5 [*KRT5*], mucin 1 [*MUC1*] and complement C3 [*C3*]) were highlighted in yellow and the prognostic ones (i.e. alpha-2-glycoprotein 1, zinc-binding [*AZGP1*], clusterin [*CLU*] and cyclin dependent kinase 1 [*CDK1*]) in pink. The colour version of this figure is available at: http://imr.sagepub.com.

than the five-gene list identified by the proposed procedure, respectively.

The interactions between the five-gene diagnostic signature and the five-gene prognostic signature were queried using the String database.[42] The corresponding network was plotted using the Cytoscape software (Figure 3).[43]

Discussion

The *KRT5* gene was identified as a discriminative gene for AC/SCC segmentation. Of note, it was selected by all bootstrapped replicates; and it has been consistently identified as a relevant gene to classify AC and SCC in previous studies even though different training sets/test sets and learning models were used.[13,32,38,39,44,45] Currently, *KRT5* is used as an immunohistochemical marker for diagnostic assays in the clinical setting.[46]

Two previous studies suggested that with only one or several genes, the AC/SCC segmentation can be successfully accomplished.[38,39] Therefore, this current study made a comparison between this three-gene list and the gene lists given by those two studies. To conclude, the current five-gene or three-gene diagnostic signature outperforms these two gene lists. Likewise, for the prognostic gene signature, a comparison between the identified five-gene list and two existing prognostic gene signatures in the literature was made.[40,41] Of note, for these two signatures some genes were outside our search space (i.e. 8023 genes under consideration) and thus had been deleted. Again, the current prognostic gene list performs the best.

Among the five genes with both high stability (frequency of being selected $> 50\%$), four of them were indicated by the GeneCards database to be directly associated with NSCLC,[37] thus being of scientific significance. For example, upregulation of *PEG10* has been reported to be associated with several malignancies such as hepatocellular carcinoma[47] and B-cell lymphocytic leukaemia.[48] A recent study concluded that the expression levels of *PEG10* were highly correlated with the TNM staging and survival time of the patients with lung cancer.[49] In Figure 3, the interactions between the five-gene diagnostic signature and the five-gene prognostic signature are given.[42,43] Specifically, four of these 10 genes were isolated (thus being deleted). In the network plot, there exist three sub-networks, including two extremely-small-sized ones in which all nodes are internal to the identified signatures and one small-sized network in which all nodes are looped together with only *CDK1* being a prognostic gene. Overall, the identified genes are independent factors to associate with their respective outcomes.

In conclusion, this current study used the GeneRank method to generate weighted gene expression profiles directly. Upon the resulting weighted expression values, the LASSO method was used to identify the relevant genes for the AC/SCC subtype segmentation and for the prognosis of NSCLC patients. The results showed that both diagnostic and prognostic gene signatures identified by the proposed procedure had a satisfactory performance, good stability, small size and meaningful biological interpretation. The identified prognostic gene signature might provide a very promising kit to predict the risk of death for AC and SCC patients as a whole. Nevertheless, a large-sized prospective study is warranted to further investigate the clinical practicability of this five-gene signature.

Availability of data and materials

The data used here were downloaded from either the GEO database or the TCGA database; both databases are open to the public.

Declaration of conflicting interest

The authors declare that there are no conflicts of interest.

Funding

The present study was supported by the Education Department of Jilin Province (No: JJKH20190032KJ) and the National Natural Science Foundation of China (No. 31401123).

ORCID iD

Suyan Tian https://orcid.org/0000-0002-5942-1542

References

1. Lemjabbar-Alaoui H, Hassan OU, Yang YW, et al. Lung cancer: biology and treatment options. *Biochim Biophys Acta* 2015; 1856: 189–210.
2. Johnson BE and Jänne PA. Basic treatment considerations using chemotherapy for patients with small cell lung cancer. *Hematol Oncol Clin North Am* 2004; 18: 309–322.
3. Simon M, Argiris A and Murren JR. Progress in the therapy of small cell lung cancer. *Crit Rev Oncol Hematol* 2004; 49: 119–133.
4. Zhu D, Ma T, Niu Z, et al. Prognostic significance of metabolic parameters measured by (18)F-fluorodeoxyglucose positron emission tomography/computed tomography in patients with small cell lung cancer. *Lung Cancer* 2011; 73: 332–337.
5. Simon G, Ginsberg RJ and Ruckdeschel JC. Small-cell lung cancer. *Chest Surg Clin N Am* 2001; 11: 165–188.
6. Yang P, Allen MS, Aubry MC, et al. Clinical features of 5,628 primary lung cancer patients: experience at Mayo Clinic from 1997 to 2003. *Chest* 2005; 128: 452–462.
7. Selvaggi G and Scagliotti G V. Histologic subtype in NSCLC: does it matter? *Oncology (Williston Park)* 2009; 23: 1133–1140.
8. Itadani H, Mizuarai S and Kotani H. Can systems biology understand pathway activation? Gene expression signatures as surrogate markers for understanding the complexity of pathway activation. *Curr Genomics* 2008; 9: 349–360.
9. Liu J, Campen A, Huang S, et al. Identification of a gene signature in cell cycle pathway for breast cancer prognosis using gene expression profiling data. *BMC Med Genomics* 2008; 1: 39.
10. Matsusue K, Kusakabe T, Noguchi T, et al. Hepatic steatosis in leptin-deficient mice is promoted by the PPARgamma target gene Fsp27. *Cell Metab* 2008; 7: 302–311.
11. Michiels S, Ternès N and Rotolo F. Statistical controversies in clinical research: prognostic gene signatures are not (yet) useful in clinical practice. *Ann Oncol* 2016; 27: 2160–2167.
12. Girard L, Rodriguez-Canales J, Behrens C, et al. An expression signature as an aid to the histologic classification of non-small cell lung cancer. *Clin Cancer Res* 2016; 22: 4880–4889.
13. Zhang A and Tian S. Classification of early-stage non-small cell lung cancer by weighing gene expression profiles with connectivity information. *Biom J* 2018; 60: 537–546.
14. Wang J, Song J, Gao Z, et al. Analysis of gene expression profiles of non-small cell lung cancer at different stages reveals significantly altered biological functions and candidate genes. *Oncol Rep* 2017; 37: 1736–1746.
15. Peng F, Wang R, Zhang Y, et al. Differential expression analysis at the individual level reveals a lncRNA prognostic signature for lung adenocarcinoma. *Mol Cancer* 2017; 16: 98.
16. Chen HY, Yu SL, Chen CH, et al. A five-gene signature and clinical outcome in non-small-cell lung cancer. *N Engl J Med* 2007; 356: 11–20.
17. Li B, Cui Y, Diehn M, et al. Development and validation of an individualized immune prognostic signature in early-stage nonsquamous non–small cell lung cancer. *JAMA Oncol* 2017; 3: 1529–1537.
18. Wistuba II, Behrens C, Lombardi F, et al. Validation of a proliferation-based expression signature as prognostic marker in early stage lung adenocarcinoma. *Clin Cancer Res* 2013; 19: 6261–6271.
19. He R and Zuo S. A robust 8-gene prognostic signature for early-stage non-small cell lung cancer. *Front Oncol* 2019; 9: 693.
20. Lin T, Fu Y, Zhang X, et al. A seven-long noncoding RNA signature predicts overall survival for patients with early stage non-small cell lung cancer. *Aging (Albany NY)* 2018; 10: 2356–2366.
21. Roepman P, Jassem J, Smit EF, et al. An immune response enriched 72-gene prognostic profile for early-stage non-small-cell lung cancer. *Clin Cancer Res* 2009; 15: 284–290.

22. Tian S, Chang HH and Wang C. Weighted-SAMGSR: combining significance analysis of microarray-gene set reduction algorithm with pathway topology-based weights to select relevant genes. *Biol Direct* 2016; 11: 50.
23. Tian S, Wang C and Wang B. Incorporating pathway information into feature selection towards better performed gene signatures. *Biomed Res Int* 2019; 2019: 2497509.
24. Kim S, Kon M and DeLisi C. Pathway-based classification of cancer subtypes. *Biol Direct* 2012; 7: 21.
25. Sokolov A, Carlin DE, Paull EO, et al. Pathway-based genomics prediction using generalized elastic net. *PLoS Comput Biol* 2016; 12: e1004790.
26. Chen L, Xuan J, Riggins RB, et al. Identifying cancer biomarkers by network-constrained support vector machines. *BMC Syst Biol* 2011; 5: 161.
27. Zhu CQ and Tsao MS. Prognostic markers in lung cancer: is it ready for prime time? *Transl Lung Cancer Res* 2014; 3: 149–158.
28. Kourou K, Exarchos TP, Exarchos KP, et al. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J* 2014; 13: 8–17.
29. Morrison JL, Breitling R, Higham DJ, et al. GeneRank: using search engine technology for the analysis of microarray experiments. *BMC Bioinformatics* 2005; 6: 233.
30. Page L, Brin S, Motwani R, et al. The pagerank citation ranking: bringing order to the web. *World Wide Web Internet Web Inf Syst* 1998; 54: 1–17.
31. Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc Ser B* 1996; 58: 267–288.
32. Tian S. Identification of subtype-specific prognostic signatures using Cox models with redundant gene elimination. *Oncol Lett* 2018; 15: 8545–8555.
33. Keshava Prasad TS, Goel R, Kandasamy K, et al. Human protein reference database – 2009 update. *Nucleic Acids Res* 2009; 37: D767–D772.
34. Friedman J, Hastie T and Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 2010; 33: 1–22.
35. Uno H, Cai T, Pencina MJ, et al. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med* 2011; 30: 1105–1117.
36. Breiman L. Bagging predictors. *Mach Learn* 1996; 24: 123–140.
37. Safran M, Dalah I, Alexander J, et al. Genecards version 3: the human gene integrator. *Database (Oxford)* 2010; 2010: baq020.
38. Zhang A, Wang C, Wang S, et al. Visualization-aided classification ensembles discriminate lung adenocarcinoma and squamous cell carcinoma samples using their gene expression profiles. *PLoS One* 2014; 9: e110052.
39. Ben-Hamo R, Boue S, Martin F, et al. Classification of lung adenocarcinoma and squamous cell carcinoma samples based on their gene expression profile in the sbv IMPROVER diagnostic signature challenge. *Syst Biomed* 2013; 1: 268–277.
40. Guo NL, Wan YW, Bose S, et al. A novel network model identified a 13-gene lung cancer prognostic signature. *Int J Comput Biol Drug Des* 2011; 4: 19–39.
41. Zhu CQ, Ding K, Strumpf D, et al. Prognostic and predictive gene signature for adjuvant chemotherapy in resected non-small-cell lung cancer. *J Clin Oncol* 2010; 28: 4417–4424.
42. Franceschini A, Szklarczyk D, Frankild S, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* 2013; 41: D808–D815.
43. Smoot ME, Ono K, Ruscheinski J, et al. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 2011; 27: 431–432.
44. Tian S and Suárez-Fariñas M. Hierarchical-TGDR: combining biological hierarchy with a regularization method for multi-class classification of lung cancer samples via high-throughput gene-expression data. *Syst Biomed* 2013; 1: 278–287.
45. Mramor M, Leban G, Demšar J, et al. Visualization-based cancer microarray data classification analysis. *Bioinformatics* 2007; 23: 2147–2154.

46. Charkiewicz R, Niklinski J, Claesen J, et al. Gene expression signature differentiates histology but not progression status of early-stage NSCLC. *Transl Oncol* 2017; 10: 450–458.

47. Okabe H, Satoh S, Furukawa Y, et al. Involvement of PEG10 in human hepatocellular carcinogenesis through interaction with SIAH1. *Cancer Res* 2003; 63: 3043–3048.

48. Hu C, Xiong J, Zhang L, et al. PEG10 activation by co-stimulation of CXCR5 and CCR7 essentially contributes to resistance to apoptosis in CD19+CD34+ B cells from patients with B cell lineage acute and chronic lymphocytic leukemia. *Cell Mol Immunol* 2004; 1: 280–294.

49. Deng X, Hu Y, Ding Q, et al. PEG10 plays a crucial role in human lung cancer proliferation, progression, prognosis and metastasis. *Oncol Rep* 2014; 32: 2159–2167.