

Identification of cancer risk assessment signature in patients with chronic obstructive pulmonary disease and exploration of the potential key genes

Qingzhou Guan^{a,b}, Peng Zhao^{a,b}, Yange Tian^{a,b}, Liping Yang^c, Zhenzhen Zhang^{a,b} and Jiansheng Li^{b,d}

^aAcademy of Chinese Medical Sciences, Henan University of Chinese Medicine, Zhengzhou, China; ^bHenan Key Laboratory of Chinese Medicine for Respiratory Disease, Co-Construction Collaborative Innovation Center for Chinese Medicine and Respiratory Diseases by Henan & Education Ministry of P.R. China, Henan University of Chinese Medicine, Zhengzhou, China; ^cSchool of Basic Medicine, Henan University of Chinese Medicine, Zhengzhou, China; ^dThe First Affiliated Hospital, Henan University of Chinese Medicine, Zhengzhou, China

ABSTRACT

It is essential to assess the cancer risk for patients with chronic obstructive pulmonary disease (COPD). Comparing gene expression data from patients with lung cancer (a total of 506 samples) and those with cancer-adjacent normal lung tissues (a total of 370 samples), we generated a qualitative transcriptional signature consisting of 2046 gene pairs. The signature was verified in an evaluation dataset comprising 18 subjects with severe disease and 52 subjects with moderate disease (Wilcoxon rank-sum test; $p = 7.33 \times 10^{-5}$). Similar results were obtained in other independent datasets. Among the gene pairs in the signature, 326 COPD stage-related gene pairs were identified based on Spearman's rank correlation tests and those gene pairs comprised 368 unique genes. Of these 368 genes, 16 genes were significantly dysregulated in COPD rat model data compared with control data. Some of these genes (*Dhx16*, *Upf2*, *Notch3*, *Sec61a1*, *Dyrk2*, and *Hmmr*) were altered when the COPD rat model was treated with traditional Chinese medicines (TCM), including Bufeiyishen formula, Bufeijianpi formula, and Yiqizishen formula. Overall, the signature could predict the cancer incidence-risk of COPD and the identified key genes might provide guidance regarding both the treatment of COPD using TCM and the prevention of cancer in patients with COPD.

KEY MESSAGES

- A cancer risk assessment signature was identified in patients with COPD.
- The signature is insensitive to batch effects and is well verified.
- COPD key genes identified in this study might play a crucial role in TCM treatment and cancer prevention.

ARTICLE HISTORY

Received 25 May 2022
Revised 25 July 2022
Accepted 6 August 2022



KEYWORDS

Chronic obstructive pulmonary disease; lung cancer; qualitative transcriptional characteristics; incidence-risk score; traditional Chinese medicines

1. Introduction

Chronic obstructive pulmonary disease (COPD) is a common respiratory disease, characterized by airflow limitation and is incompletely reversible [1,2]. Patients with COPD suffer decline in lung function resulting in a severe compromise in the quality of life and imposing heavy economic burdens on patients, families, and society [3,4]. The overall incidence of COPD was reported to be 8.6% in China, and was as high as 13.7% for individuals aged 40 years or older [5]. In 2019, the global prevalence of COPD among people aged 30–79 years was 10.3% (95% CI 8.2–12.8) using

the GOLD case definition, which translates to 391.9 million people (95% CI 312.6–487.9) [6]. Moreover, COPD is an independent high-risk factor for the occurrence of lung cancer [7,8]. Lung cancer could develop from COPD through a continuous, multi-step process whereby normal lungs advance to moderate and then severe COPD, and eventually develop into cancer [9]. However, to the best of our knowledge, there is currently no molecular signature to accurately assess the risk of cancer incidence among patients with COPD. Thus, there is significant clinical value in developing a molecular signature for assessing the incidence of COPD converting to lung cancer. Traditional Chinese

CONTACT Jiansheng Li  li_js8@163.com  Henan Key Laboratory of Chinese Medicine for Respiratory Disease, Co-Construction Collaborative Innovation Center for Chinese Medicine and Respiratory Diseases by Henan & Education Ministry of P.R. China, Henan University of Chinese Medicine, Zhengzhou 450046, China

 Supplemental data for this article can be accessed at <https://doi.org/10.1080/07853890.2022.2112070>.

© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

medicines (TCM) have unique merits, exhibiting high efficacy and fewer adverse reactions, and some TCM have been successfully applied for the treatment of COPD in clinical settings [10–12]. For example, clinical studies have revealed Tiaobu Feishen formulae (TBFS), including Bufeif Yishen formula (BYF), Bufeif Jianpi formula (BJF), and Yiqi Zishen formula (YZF), had desirable pharmacological effects on COPD, such as alleviating the clinical symptoms of patients with stable COPD, reducing the exacerbation frequency, delaying acute exacerbation, and improving pulmonary function and exercise capacity [13]. Moreover, these three formulae have demonstrated beneficial effects in COPD rat model, inhibiting expression of inflammatory cytokines, protease–antiprotease imbalance, and collagen deposition [10,14–17]. Among the genes constituting the cancer risk signature, it will also be of significance to identify potential key genes that are reversed when the COPD rat model is treated with those TCMS, which might guide COPD treatment using TCM and aid in the prevention of lung cancer occurrence from COPD.

High-throughput gene detection technology has become widely applied, and various quantitative transcriptional signatures have been used in subtyping diseases and early diagnosis [8,18–21]. Nevertheless, due to batch effects, these types of signatures are not suitable for the analysis of individuals and are therefore difficult to apply them in clinical practice. Several disease signatures based on quantitative transcriptional feature, such as AlloMap[®] [22], have already been approved by the US Food and Drug Administration (FDA). However, because of batch effects, those samples must be measured in specific laboratories, which also limits their clinical application. Qualitative transcriptional characteristics, also called within individual sample relative expression orderings (REOs) of genes, are robust solutions to the batch effect problem and suitable for individualized analysis in clinical practice [22,23]. Using the robust performance of qualitative transcriptional characteristics, researchers can merge data detected by the same or similar platforms from multiple sources to train classifier models or signatures, which would easily obtain robust signatures [2,24,25]. Furthermore, the technique is suitable for samples detected using different platforms.

Based on the unique merits of qualitative transcriptional characteristics, this study identified a cancer incidence-risk signature for patients with COPD without cancer, and the performance of the signature was verified in multiple independent datasets.

Table 1. Data analysed in this study.

GEO No.	Gene ^a	Platform	Normal sample size	Cancer sample size
GSE19804	20486	Affymetrix GPL570	60	60
GSE18842	20486	Affymetrix GPL570	45	46
GSE27262	20486	Affymetrix GPL570	25	25
GSE31210	20486	Affymetrix GPL570	20	226
GSE19188	20486	Affymetrix GPL570	65	91
GSE32863	25186	Illumina GPL6884	58	58
GSE31267	24384	Illumina GPL6947	24	–
GSE15197	18615	Agilent GPL6480	13	–
GSE40588	19595	Agilent GPL6480	60	–

^aThe number of genes detected in the corresponding dataset.
–: there is no sample in the corresponding category.

Furthermore, among the genes constituting the cancer risk signature, COPD key genes that could be regulated by TCM were identified, and the value of these COPD key genes in drug treatment and cancer prevention warrants further exploration.

2. Materials and methods

2.1. Public data and preprocessing

Gene expression profiles of lung cancer and normal lung tissue samples from multiple sources were downloaded from the GEO database (Table 1). For data detected by the Affymetrix platform, the raw mRNA expression data (.CEL files) was downloaded and the Robust Multi-array Average (RMA) algorithm was applied for preprocessing. For data detected by Illumina or Agilent platforms, the processed data were directly downloaded. All cancer samples were from surgical resection in patients with non-small cell lung carcinoma (NSCLC), while the normal samples were obtained from adjacent normal tissues of patients with lung cancer.

For the downloaded data, when multiple probes mapped to an identical gene, the measurement of the gene was calculated as the arithmetic mean value of the multiple probe values. When a probe mapped to none or more than one gene, the probe data were discarded.

2.2. COPD rat data and drug treatment

The rat data analysed in this study were obtained from our previous study [10–12,26] and a COPD model generated using Sprague–Dawley rats was prepared as previously described [27]. Briefly, the rats were exposed to cigarette smoke and repeated *Klebsiella pneumoniae* infections. In the ninth week, COPD model rats were randomly divided into five groups as shown in Supplementary Table S1. The groups of rats were intragastrically treated with normal saline (model

group, 2 mL/animal), aminophylline (APL, 2.3 mg/kg), BYF, BJJ, or YZF each day from weeks 9 to 20, respectively; the drug concentrations and dose of the three TCMs are shown in [Supplementary Table S2](#). Dosages of the TCM formulae were calculated according to the clinically used dosages of adult patients and the body surface area conversion equation between human and rat: $D_{\text{rat}}=D_{\text{human}}\times(l_{\text{rat}}/l_{\text{human}})\times(W_{\text{rat}}/W_{\text{human}})^{2/3}$, where D is dose, l is body shape index, and W is body weight. The control group rats were fed with normal saline intragastrically (2 mL). Each group included six replicates and the rats in each group were separately given the corresponding drug or normal saline treatment. All animals were handled humanely during the process of the experiment and were anaesthetized and sacrificed to obtain lung tissues on week 32. The components of BYF, BJJ, YZF, and APL were described in previous studies [10–12]. Mass spectrometry and high-performance liquid chromatography fingerprint were respectively performed in previous studies to identify the main chemical constituents of BYF and BJJ [28,29].

Briefly, BYF (patent: ZL.201110117578.1) is composed of 12 Chinese medicinal herbs, including *Panax ginseng* C.A.Mey. 9g, *Astragalus mongholicus* Bunge 15g, *Cornus officinalis* Siebold & Zucc. 12g, *Lycium barbarum* L. 12g, *Schisandra chinensis* (Turcz.) Baill. 9g, *Epimedium sagittatum* (Siebold & Zucc.) Maxim. 9g, *Fritillaria thunbergii* Miq. 9g, *Paeonia lactiflora* Pall. 9g, *Pheretima* 12g, *Perilla frutescens* (L.) Britton 9g, *Ardisia japonica* (Thunb.) Blume 15g, and *Citrus × aurantium* L 9g, which were also reported in our previous studies [11,30]. Similarly, the components of BJJ also included 12 Chinese medicines: *Astragalus mongholicus* Bunge 15g, *Polygonatum sibiricum* Redouté 15g, *Codonopsis pilosula* (Franch.) Nannf. 15g, *Atractylodes macrocephala* Koidz. 12g, *Poria cocos* (Schw.) wolf 12g, *Fritillaria thunbergii* Miq. 9g, *Pheretima* 12g, *Magnolia officinalis* Rehder & E.H.Wilson 9g, *Citrus × aurantium* L. 9g, *Aster tataricus* L.f. 9g, *Ardisia japonica* (Thunb.) Blume 15g, and *Epimedium sagittatum* (Siebold & Zucc.) Maxim 6g, as also shown in one of our previous studies [10]. YZF is composed of 13 Chinese medicines, including *Panax ginseng* C.A.Mey. 9g, *Polygonatum sibiricum* Redouté 15g, *Ophiopogon japonicus* (Thunb.) Ker Gawl. 15g, *Schisandra chinensis* (Turcz.) Baill. 9g, *Lycium barbarum* L. 12g, *Rehmannia glutinosa* (Gaertn.) DC. 15g, *Neolitsea cassia* (L.) Kosterm. 3g, *Fritillaria thunbergii* Miq. 9g, *Pheretima* 12g, *Paeonia × suffruticosa* Andrews 12g, *Perilla frutescens* (L.) Britton 9g, *Stemona tuberosa*

Lour. 9g, and *Citrus × aurantium* L 9g [12]. Plant names were verified according to the Kew search tool. However, due to *Pheretima* and *Poria cocos* (Schw.) wolf not belonging to the scope of botanical medicinal materials, they were verified by searching literature and “Chinese Pharmacopoeia”. APL was obtained from Shandong Xinhua Pharmaceutical Co., Ltd. (Shandong, China). *K. pneumoniae* (strain ID: 46114) was purchased from the National Centre for Medical Culture Collection (CMCC, Beijing, China). The herbs were identified and prepared in fluid extract [10–12]. This study was approved by the Experimental Animal Care and Ethics Committee of the First Affiliated Hospital, Henan University of Chinese Medicine (2012HLD-0001).

For the six replicate samples from each group, RNA was extracted and purified from lung tissues using TRIzol reagent and Qiagen RNeasy Micro Kit, and then was measured by Agilent Whole Rat Genome Oligo Microarray. Raw data obtained in the above process were preprocessed with Agilent GeneSpring GX software (version 11.0). Differential expression analysis between two of these groups was performed using Student’s t -tests. In this present study, a cancer risk assessment signature for patients with COPD was firstly identified, and its reliability was verified in independent data. Among the genes constituting the cancer risk signature, the previously produced gene expression data of rat were applied to identify COPD key genes and further identify the genes that were reversed after drug treatment. The resulting information could potentially provide some guidance regarding the treatment of COPD and the prevention of cancer.

2.3. Identification of the qualitative transcriptional signature

Between the gene expression data of lung cancer and normal lung tissues from the training set (as shown in [Table 1](#)), highly stable gene pairs with opposite REOs were identified as the signature to predict the cancer incidence-risk in patients with COPD (with a threshold of 90%).

For the genes detected in a specific type of tissue sample from training datasets, all genes were pairwise compared to select stable gene pairs. For two genes, such as gene A and gene B, in one sample, their REO pattern was identified as $A > B$ (or $A < B$) if the measurement of gene A was larger (or smaller) compared with that of gene B. In this study, a gene pair was considered highly stable when the gene pair (A, B)

had an identical REO pattern in at least 90% of samples. Among the samples from two groups, if one gene pair was stable in both of groups but with reversal REO pattern, this gene pair was considered a reversal gene pair. Finally, from all gene pair combinations, the reversal gene pairs were selected and this was considered the signature for predicting the cancer incidence-risk.

For the gene pairs contained in the identified signature, the REO pattern of gene pairs representing lung cancer was used to calculate the cancer risk score in patients with COPD. For each patient, the cancer incidence-risk score was defined as the percentage of gene pairs characterizing lung cancer among the gene pairs of the signature. Supposing the number of gene pairs of the signature was m , among which n gene pairs had REO patterns characterizing lung cancer in this particular sample, then the incidence-risk score is given by n/m . The property of the identified signature was then verified in samples of non-cancer patients with COPD at different disease courses from several datasets.

2.4. KEGG pathway enrichment

A total of 330 KEGG pathways consisting of 7838 genes were obtained from Kyoto Encyclopaedia of Genes and Genomes (KEGG) database [31]. The significance of the pathways was determined by hypergeometric distribution model, calculated as the following:

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{m}{i} \binom{N-m}{n-i}}{\binom{N}{n}}$$

where m indicates the number of genes annotated in one given pathway, n indicates the number of interested genes, N indicates the number of genes detected by the high-throughput platform, and k indicates the number of interested genes in the given pathway.

3. Results

3.1. Identification of a molecular signature to build the cancer incidence-risk score

The flowchart of this study is shown in Figure 1. Considering that lung cancer develops in a continuous, multistep process from normal lung tissues, with the threshold of 90%, stable gene pairs with opposite REOs between lung cancer and normal lung tissue samples were identified (see Materials and Methods).

For the 448 lung cancer and 215 normal lung tissue samples obtained from the five datasets detected by Affymetrix platform (as shown in Table 1), with the threshold of 90%, 21,612 stable gene pairs in both lung cancer and normal lung tissues but with reversed REO patterns were obtained; these gene pairs were considered stable reversal gene pairs. For the 58 lung cancer and 82 normal lung tissue samples obtained from the two datasets detected by Illumina platform (as shown in Table 1), 337,728 stable reversal gene pairs were obtained with the same threshold (90%). Among the two lists of stable reversal gene pairs obtained above, 3716 gene pairs were consistently identified. Based on those 3716 gene pairs, there were 2046 gene pairs with same REO patterns in more than 90% of the 73 normal lung tissues data detected by the Agilent platform. These 2046 gene pairs (Supplementary Table S3), including 1700 unique genes (Supplementary Table S4), were identified as the molecular signature and the percentage of the gene pairs characterizing lung cancer tissues were applied to predict the cancer incidence-risk score of non-cancer patients with COPD (see Materials and Methods). For a total of 506 lung cancer and 370 normal lung tissue samples in the training data, based on our signature consisting of 2046 gene pairs, the area under the receiver operating characteristic curve (AUC) value was 0.9929 (95% CI, 0.9634–1) (Supplementary Figure S1). The performance of the signature was then evaluated among patients with COPD from multiple datasets by comparing cancer risk scores of patients with COPD at different disease courses.

Additionally, measurements of the genes **BIRC5** and **ASPA**, **BARD1** and **PTPRB**, **CCNA2** and **ACKR4** in lung cancer and normal lung tissue samples from datasets GSE18842 and GSE27262 were taken as an example to show that the qualitative transcriptional characteristics are robust in normal tissue samples (the expression value of **ASPA** (**PTPRB**, **ACKR4**) is higher compared with that of **BIRC5** (**BARD1**, **CCNA2**)) but reversed in cancer tissues (Figure 2). This would provide a basis for the selection of the cancer risk signature for non-cancer patients with COPD.

3.2. Performance of the signature in COPD samples at different disease courses

The pathophysiological process of lung cancer involves transforming normal lung to lungs affected by COPD, culminating in outright malignant transformation [9]. The performance of the signature (whose score ranges from 0 to 1) was therefore evaluated in

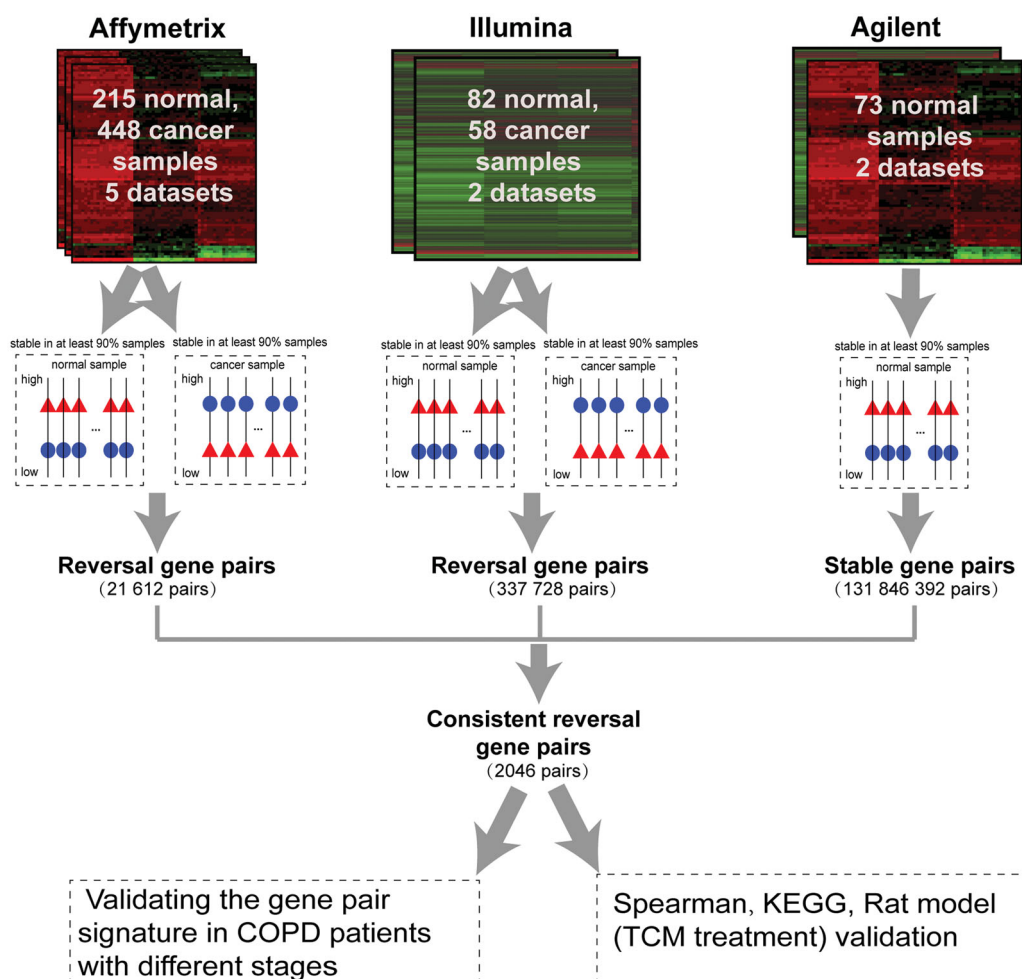


Figure 1. Analysis flowchart for this study.

COPD samples with different disease courses. Higher risk scores correlated with a greater cancer risk.

In the dataset GSE69818, including 18 severe and 52 moderate COPD samples, the median of cancer incidence-risk score in severe COPD data was 0.0864, significantly higher than that in the moderate COPD samples (Wilcoxon rank-sum test; $p = 7.33 \times 10^{-5}$). In the datasets GSE76925, containing 111 severe COPD samples, and GSE37768, comprising 18 moderate COPD samples, similar results (Wilcoxon rank-sum test; $p = 1.67 \times 10^{-8}$) were obtained (Figure 3 and Supplementary Table S5). Moreover, the risk scores in samples from 18 patients with severe COPD that came from dataset GSE69818 were also significantly higher compared with those of the 18 moderate COPD samples from dataset GSE37768 (Wilcoxon rank-sum test; $p = 2.39 \times 10^{-5}$). Similar results were obtained in the analysis of severe and moderate COPD samples from datasets GSE76925 and GSE69818 (Wilcoxon rank-sum test; $p = 5.40 \times 10^{-5}$). These data suggest that our

signature could be applied to various samples from multiple sources, highlighting the cross-platform performance of the signature.

3.3. KEGG pathway enrichment analysis based on signature genes

For the 2046 gene pairs in the signature, COPD stage-related gene pairs were further identified based on Spearman's rank correlation tests. For the dataset GSE69818, consisting of 53 moderate and 18 severe COPD samples, 33 stage-related gene pairs were identified with false discovery rate (FDR) $< 5\%$. Similarly, for the combined data of GSE76925 and GSE37768, 301 stage-related gene pairs were identified. For the above two lists of gene pairs, eight gene pairs were commonly identified and this was statistically significant (hypergeometric distribution model, $p = 9.97 \times 10^{-2}$). The two lists of gene pairs were then combined as the COPD stage-related gene pairs (a

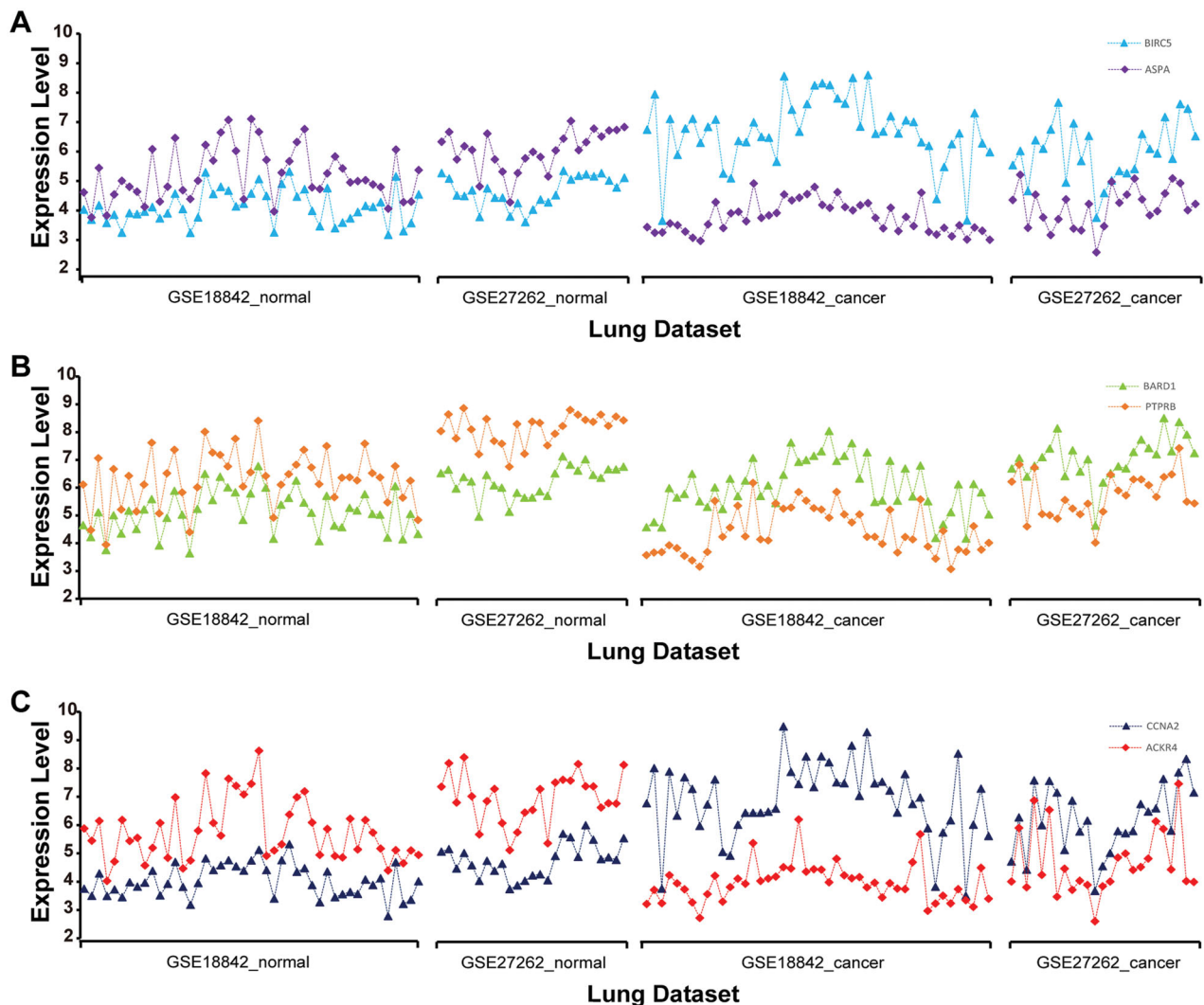


Figure 2. Distribution of gene expression levels for the three gene pairs—*BIRC5-ASPA* (A), *BARD1-PTRPB* (B), and *CCNA2-ACKR4* (C)—in GSE18842 and GSE27262 datasets. Horizontal coordinates represent cancer and normal lung tissue from datasets GSE18842 and GSE27262. Vertical coordinates represent the expression level of the corresponding gene.

total of 326 gene pairs) for subsequent analysis, and 368 unique genes were included among those gene pairs. Based on these 368 genes and the hypergeometric distribution model, KEGG pathway analysis was performed. With $FDR < 5\%$, no significantly correlated pathway was enriched, which might be ascribed to insufficient statistical power due to too few genes of interest. Therefore, pathway enrichment analysis was also performed under a relatively loose threshold condition. With $p < 5\%$, six significantly related pathways were enriched (Supplementary Table S6) and these pathways are related to the progress of COPD. For instance, studies showed that there was 2.5-fold in COPD samples compared with the normal control samples for RNA polymerase II occupancy at the promoter [32]. CoQ10 or ubiquinone levels were decreased in patients with COPD, probably due to the defense response of the organism [33,34]. Beta-

adrenoceptor-mediated lipolysis and thermogenesis are impaired in patients with COPD [35].

3.4. COPD-related genes in the rat model

The 368 human COPD stage-related genes identified above were ortholog converted to rat genes using the biological DataBase network [36], and 340 genes were obtained. Of these, only the measurements of 16 genes were significantly altered in the COPD rat model compared with control groups (six vs six) (Table 2).

Among these 16 genes, those genes that were reversed when undergoing drug treatment (with BYF, BJB, YZF, and APL) were subsequently identified (Supplementary Tables S7–S10). The frequency of genes that were reversed after using this treatment protocol was then calculated (Table 3). With a

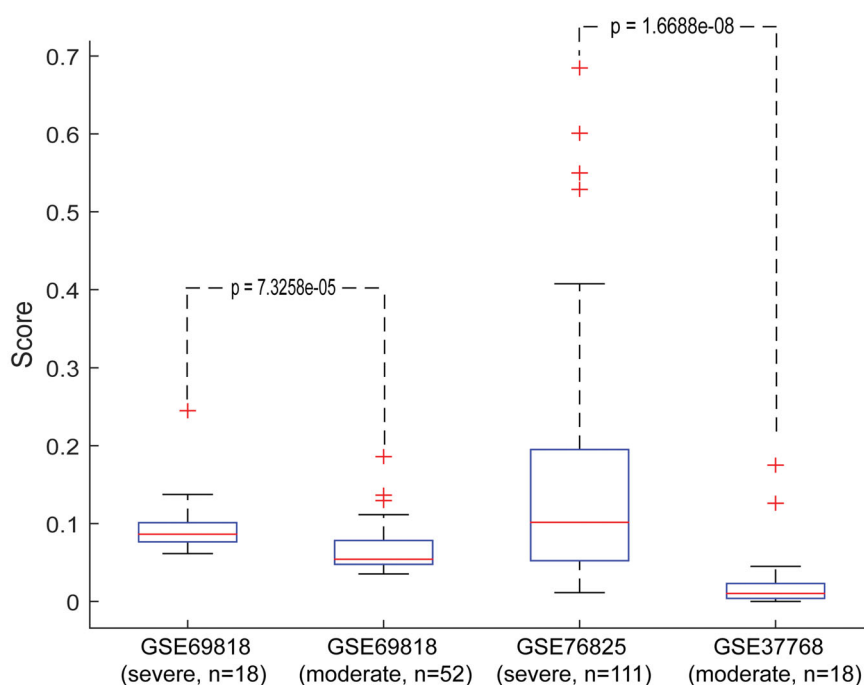


Figure 3. Performance of the signature in COPD samples with different disease courses. Horizontal coordinates represent severe and moderate COPD samples from public database. Vertical coordinates represent the score of our signature in severe and moderate COPD samples. The Wilcoxon rank-sum test was applied to calculate the p values.

Table 2. Differentially expressed genes between COPD rat model and control group.

Gene symbol	FC	T	p
<i>Dhx16</i>	0.530623	-3.72567	.003938
<i>Upf2</i>	1.039416	3.475603	.005965
<i>Uqcrc2</i>	1.015124	3.212685	.00929
<i>Rhobtb3</i>	0.884189	-3.09422	.011362
<i>Znhit3</i>	1.012759	3.046952	.012316
<i>Denr</i>	1.013586	3.017057	.01296
<i>Notch3</i>	0.975045	-2.97443	.01394
<i>Sdhc</i>	0.984546	-2.5892	.026987
<i>Dyrk2</i>	0.923138	-2.53068	.029836
<i>Sec61a1</i>	1.039209	2.52876	.029934
<i>Hmmr</i>	0.688459	-2.48504	.032263
<i>Exosc7</i>	0.943282	-2.48184	.032441
<i>Sh3gl3</i>	0.718159	-2.42208	.035933
<i>Noa1</i>	0.971469	-2.28776	.045186
<i>Trrap</i>	1.043465	2.252802	.047951
<i>Plk4</i>	0.857672	-2.24158	.048873

FC: fold-change of the COPD rat samples compared with control samples; T : test statistic value between COPD rat and control samples using the Student's t -tests.

relatively loose threshold ($p < .2$), the genes *Dhx16*, *Upf2*, *Denr*, *Notch3*, *Dyrk2*, *Sec61a1*, *Hmmr*, and *Noa1* were reversed in at least three treatment protocols and most of these genes were reported to be related to COPD or lung cancer [37–41]. The value of these genes warrants further study in the future.

4. Discussion

Using qualitative transcriptional features, a signature for the cancer incidence-risk assessment of non-cancer

patients with COPD was identified. The signature was subsequently validated in patients with COPD at different disease courses from multiple data sources. This method was successfully applied in a previous study for assessing colorectal cancer incidence-risk among patients with precancerous colorectal lesions [42]. Carcinogenesis of lung cancer is a continuous, multi-step malignant transformation process from normal lung tissues. One of pathogenic types of lung cancer arises from normal lung tissues advancing to moderate and then severe COPD, and eventually developing into cancer. The signature in the current study was developed based on normal lung and lung cancer tissue samples. Thus, the genes constituting the cancer risk signature might play vital roles in lung cancer or COPD pathogenesis. Based on the signature genes, key genes of COPD were further identified by correlation analysis and further optimized in control rat data and COPD rat model data with and without TCM treatment, which might guide efforts for cancer prevention and the treatment of COPD by TCM.

Most of the genes reproduced in the rat model were reported to be related to lung cancer or/and COPD. For example, dysregulation of *Notch1* and *Notch3* has recently been reported to be correlated with the pathogenesis of COPD [37]. The *Notch3* downstream target *HEYL* is an important regulator of airway epithelial cell proliferation and differentiation. Reduced expression of *HEYL* correlates with the

Table 3. Frequency of genes that were reversed between the treatment and model group.

Gene symbol	FC (M_vs_C)	T (M_vs_C)	p_value (M_vs_C)	p_value (BYF_vs_M)	p_value (BJF_vs_M)	p_value (YZF_vs_M)	p_value (APL_vs_M)	Num $p < .05$	Num $p < .1$	Num $p < .2$
<i>Dhx16</i>	0.530623	-3.72567	.003938	0.701224	5.52E-09	0.097164	5.50E-05	2	3	3
<i>Upf2</i>	1.039416	3.475603	.005965	.027316	.011337	.058669	.02064	3	4	4
<i>Uqcrc2</i>	1.015124	3.212685	.00929	.022126	.043653	.93704	.822763	2	2	2
<i>Rhobtb3</i>	0.884189	-3.09422	.011362	.78651	.197268	.106184	.211585	0	0	2
<i>Znhit3</i>	1.012759	3.046952	.012316	.560379	.689214	.271842	.713968	0	0	0
<i>Denr</i>	1.013586	3.017057	.01296	.010532	.020327	.118537	.100091	2	2	4
<i>Notch3</i>	0.975045	-2.97443	.01394	.475693	.000212	.088364	.020111	2	3	3
<i>Sdhc</i>	0.984546	-2.5892	.026987	.728516	.001616	.317634	.000346	2	2	2
<i>Dyrk2</i>	0.923138	-2.53068	.029836	.718672	.000155	.19656	.000206	2	2	3
<i>Sec61a1</i>	1.039209	2.52876	.029934	.202428	.056216	.006779	.014547	2	3	3
<i>Hmmr</i>	0.688459	-2.48504	.032263	.994814	.124428	.03643	.016983	2	2	3
<i>Exosc7</i>	0.943282	-2.48184	.032441	.492618	.000315	.509816	.000261	2	2	2
<i>Sh3gl3</i>	0.718159	-2.42208	.035933	.36409	.814336	.458153	.089266	0	1	1
<i>Noa1</i>	0.971469	-2.28776	.045186	.138333	.00598	.351014	.173025	1	1	3
<i>Trrap</i>	1.043465	2.252802	.047951	.656511	.036755	.571989	.012966	2	2	2
<i>Plk4</i>	0.857672	-2.24158	.048873	.547285	.018171	.724516	.011406	2	2	2

FC: fold-change of COPD rat samples compared with control samples; T: test statistic value between COPD rat and control samples using the Student's *t*-tests; p_value: *p* value between the corresponding two group samples (including Model vs Control, BYF vs Model, BJF vs Model, YZF vs Model, and APL vs Model) using Student's *t*-tests; Num: number of the corresponding genes occurring in the four treatment protocols with one certain threshold.

impaired differentiation capacity of COPD primary human bronchial epithelial cells and overexpression of *HEYL* in COPD cells promoted differentiation into club, goblet, and ciliated cells [43]. Moreover, Sun et al. [1] found *Notch3* was downregulated in patients with COPD and could be targeted by miR-206. *Notch3* was also reported to be related to lung cancer. In three NSCLC cell lines (H292, A549, and Calu-3), Shi et al. [44] proved that overexpression of *NOTCH3* was related to increased cell growth rate, migration, and invasiveness abilities, as well as decreased apoptosis rate. Furthermore, si-RNA transfection in these NSCLC cell lines reversed these cellular biological behaviours [44]. *Notch3* can promote colony formation and sphere formation of stem-like capacity in lung cancer cells, and high expression of *Notch3* was related to a poor outcome of patients with NSCLC [45]. The missense mutation rate of *UPF1* or *UPF2* was higher in lung cancer [46]. *UPF2* binds *UPF1*, one of its family proteins, with a high affinity [47]. Through interaction with *UPF1* to promote *ZFPM2* mRNA decay, *ZFPM2*-AS1 could promote lung adenocarcinoma (LUAD) cell growth, migration, and the epithelial-mesenchymal transition process, thus exerting oncogenic functions [48]. The single nucleotide polymorphism (SNP) of rs115420460 in *DHX16* was significantly different in lung cancer samples compared with controls from the TRICL Consortium, and was demonstrated to be associated with lung cancer risk. Moreover, the location of this SNP was within the previously identified lung-cancer-susceptible region Chr6p21.33 and in high linkage disequilibrium with previously reported lung cancer SNPs from genome-wide association studies [49]. *DYRK2* might play an essential role in NSCLC, and its expression may predict the chemotherapy response in

patients with NSCLC [38,39]. The expression level of *DYRK2* was significantly increased in lung cancer tissues compared with normal tissues, which might indicate a potential role of *DYRK2* in lung cancer development and/or progression [50]. Moreover, *DYRK2* was also overexpressed among lung cancer (LUAD and LUSC) in TCGA data [51]. *HMMR* is involved in lung cancer progression and is significantly associated with outcome [40,41]. *HMMR* is an independent risk factor for LUAD, and its high expression was significantly correlated with poor clinicopathological features and adverse outcomes (progression and metastasis of LAUD), whose expression may affect tumorigenic progression by altering the tumour micro-environment and playing a pivotal role in immune response regulation [52,53]. *DENR* was reported to be a risk gene in lung cancer, and its high expression could inhibit the survival of patients with lung cancer [54]. Further research on these genes might provide some valuable guidance for cancer prevention and TCM treatment of COPD.

On the other hand, our cancer risk signature in patients with COPD was developed based on normal lung and lung cancer tissue samples. Thus, the signature had the potential to discriminate lung cancer from normal lung tissues, and this ability was subsequently verified using independent data. Based on the majority vote rule, for the 59 normal lung tissues and 594 lung cancer tissues obtained from TCGA, the signature identified in the current study has excellent discriminating ability, and the values of AUC, sensitivity, and specificity were 0.9981 (95% CI, 0.6420–1), 93.64%, and 100.00%, respectively (Supplementary Figure S2). Similarly, for the 30 normal lung tissues and 36 lung cancer tissues obtained from GSE7670, the values of

AUC, sensitivity, and specificity were 0.9991 (95% CI, 0.7441–1), 94.44%, and 100%, respectively. For the 30 normal lung tissues and 80 lung cancer tissues obtained from GSE43458, the values of AUC, sensitivity, and specificity were 0.9835 (95% CI, 0.6603–1), 80.00%, and 96.67%, respectively, and for the 20 normal lung tissues and 80 lung cancer tissues obtained from GSE33532, the values were 1.000 (95% CI, 0.5839–1), 100.00%, and 100.00%, respectively. These results demonstrated that the signature has the ability to discriminate lung cancer from normal lung tissues. Moreover, the performance of the signature was also validated in COPD-only patients and COPD patients with lung cancer by searching the gene expression data of lung tissues from these two groups of patients. One dataset (GSE8581) with COPD lung tissues from COPD patients with lung cancer, and three datasets (GSE103174, GSE151052, and GSE106986) with COPD lung tissues from COPD-only patients were obtained. There was no dataset that simultaneously contained lung tissue samples from COPD-only patients and COPD patients with lung cancer. Thus, the performance of the signature to predict cancer incidence-risk of COPD patients was validated in samples from different datasets. The risk scores in 15 COPD samples from COPD patients with lung cancer from dataset GSE8581 were significantly higher compared with those in 37 samples from COPD-only patients from dataset GSE103174 (Wilcoxon rank-sum test; $p = 1.10 \times 10^{-8}$). Similarly, the risk scores in COPD samples from GSE8581 were also significantly higher compared with those in 77 samples from COPD-only patients from dataset GSE151052 (Wilcoxon rank-sum test; $p = 5.34 \times 10^{-10}$). However, the risk scores were not significantly different between the 15 COPD patients with lung cancer from dataset GSE8581 and 14 samples from COPD-only patients from dataset GSE106986 (Wilcoxon rank-sum test; $p = .68$), which might be ascribed to low statistical power due to small sample size. These results further demonstrated that the signature could effectively predict the cancer incidence-risk of patients with COPD and also exhibited cross-platform ability.

Due to the lack of corresponding clinical follow-up data, it is not possible to verify whether the individuals without cancer and with high lung cancer incidence-risk score, as identified by the signature, eventually develop into cancer. Future studies will involve collaboration with affiliated hospitals to better evaluate the robustness of the signature; patients would be followed to further appraise the robustness of the signature and to compare the cancer

incidence-risk score, calculated by the signature, with the time from diagnosis to carcinogenesis. This will determine whether individuals at high risk of cancer (calculated based on the signature) eventually develop into cancer. The financial burden of high-throughput sequencing is markedly decreasing. Consequently, for the scarce precious tissue samples at the clinical practice, it will be possible to simultaneously measure a set of disease genes that could more fully reveal the value of clinical samples under controllable cost conditions. Such data could be reused in other studies for different application scenarios involving diagnosis, histological classification, prognoses evaluation of disease, etc., thereby enhancing the value of the clinical research.

In conclusion, the molecular signature identified in this study (based on qualitative transcriptional characteristics) circumvents problems related to batch effects [55,56], variations in tumour epithelial cells from different sampling sites [57], partial RNA degradation [58], and amplification bias of minimum specimens [59]. The signature is suitable for inaccurately sampled tissues and can be applied for individualized analysis, which is more in line with the clinical setting [22]. Moreover, the reversed genes identified in the COPD rat model and drug treatment group might play a key role in medical treatment of COPD, and this warrants further investigation.

5. Conclusions

COPD is a common disease with severe health consequences. It is also a high-risk factor for lung cancer. For the non-cancer patients with COPD, it would be significant if their cancer incidence-risk could be assessed. Considering the unique merits of qualitative transcriptional characteristics (also called the within samples REOs of genes), which are insensitive to batch effects and could be used for the analysis of individual patients, a qualitative signature was identified to predict the lung cancer incidence-risk for non-cancer patients with COPD. Key genes for COPD were further identified, optimized by correlation analysis with COPD stage, and filtered in COPD rat model data. The genes that occurred in reverse fashion when the COPD rat model was treated with some TCM were further identified. In summary, the qualitative transcriptional signature circumvented problems associated with batch effects and is suitable for the individualized diagnosis of single samples, making it feasible for application in clinical settings for the surveillance of non-cancer patients with COPD. The value of COPD

key genes in both TCM treatment of COPD and cancer prevention should be further explored.

Ethical approval

All human data used in this study were obtained from the GEO public database (<https://www.ncbi.nlm.nih.gov/geo/>). Rat data analysed in this study were obtained from our previous study [10–12,26]. The study was approved by the Experimental Animal Care and Ethics Committee of the First Affiliated Hospital, Henan University of Chinese Medicine (2012HLD-0001) in 2012.

Author contributions

QZG searched and analysed data, created figures, and wrote the manuscript. PZ, YGT, and LPY constructed the COPD rat model and performed corresponding animal experiment. ZZZ searched and analysed data, and produced tables and supplementary files. JSL and QZG designed the study, revised the manuscript, and checked the work. All authors read and approved the final manuscript.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This study was supported by China National Postdoctoral Program for Innovative Talents [grant number BX20200115], the plan for the Key Scientific Research Foundation of the Higher Education Institutions of Henan Province [grant number 21A360019], National Natural Science Foundation of China [grant number 81973822], and Scientific Nursery Research Program of Henan University of Chinese Medicine [grant number MP2021-16].

Data availability statement

The data that support the findings of this study are available on request from the first or corresponding author.

References

- [1] Sun Y, An N, Li J, et al. miRNA-206 regulates human pulmonary microvascular endothelial cell apoptosis via targeting in chronic obstructive pulmonary disease. *J Cell Biochem.* 2019;120(4):6223–6236.
- [2] Xia J, Zhang H, Guan Q, et al. Qualitative diagnostic signature for pancreatic ductal adenocarcinoma based on the within-sample relative expression orderings. *J Gastroenterol Hepatol.* 2021;36(6):1714–1720.
- [3] Mao J, Li Y, Li S, et al. Bufeï Jianpi granules reduce quadriceps muscular cell apoptosis by improving mitochondrial function in rats with chronic obstructive pulmonary disease. *Evid Based Complement Alternat Med.* 2019;2019:1–9.
- [4] Ma J, Tian Y, Li J, et al. Effect of Bufeï Yishen granules combined with electroacupuncture in rats with chronic obstructive pulmonary disease via the regulation of TLR-4/NF-kappaB signaling. *Evid Based Complement Alternat Med.* 2019;2019:6708645.
- [5] Wang C, Xu J, Yang L, et al. Prevalence and risk factors of chronic obstructive pulmonary disease in China (the China Pulmonary Health [CPH] study): a national cross-sectional study. *Lancet.* 2018;391(10131):1706–1717.
- [6] Adeloje D, Song P, Zhu Y, et al. Global, regional, and national prevalence of, and risk factors for, chronic obstructive pulmonary disease (COPD) in 2019: a systematic review and modelling analysis. *Lancet Respir Med.* 2022;10(5):447–458.
- [7] Sandri BJ, Kaplan A, Hodgson SW, et al. Multi-omic molecular profiling of lung cancer in COPD. *Eur Respir J.* 2018;52(1):1702665.
- [8] Xia R, Tang H, Shen J, et al. Prognostic value of a novel glycolysis-related gene expression signature for gastrointestinal cancer in the Asian population. *Cancer Cell Int.* 2021;21(1):154.
- [9] Wang X. New biomarkers and therapeutics can be discovered during COPD-lung cancer transition. *Cell Biol Toxicol.* 2016;32(5):359–361.
- [10] Zhao P, Yang L, Li J, et al. Combining systems pharmacology, transcriptomics, and metabolomics to dissect the therapeutic mechanism of Chinese Herbal Bufeï Jianpi formula for application to COPD. *Int J Chron Obstruct Pulmon Dis.* 2016;11:553–566.
- [11] Li J, Zhao P, Yang L, et al. System biology analysis of long-term effect and mechanism of Bufeï Yishen on COPD revealed by system pharmacology and 3-omics profiling. *Sci Rep.* 2016;6:25492.
- [12] Li J, Zhao P, Tian Y, et al. Systems pharmacology-based dissection of the active ingredients and targets of Yiqi Zishen formula for application to COPD. *Int J Clin Exp Med.* 2017;10(8):12825–12844.
- [13] Li SY, Li JS, Wang MH, et al. Effects of comprehensive therapy based on traditional Chinese medicine patterns in stable chronic obstructive pulmonary disease: a four-center, open-label, randomized, controlled study. *BMC Complement Altern Med.* 2012;12:197.
- [14] Li J, Zhao P, Li Y, et al. Systems pharmacology-based dissection of mechanisms of Chinese medicinal formula Bufeï Yishen as an effective treatment for chronic obstructive pulmonary disease. *Sci Rep.* 2015;5:15290.
- [15] Zhao P, Li J, Li Y, et al. Systems pharmacology-based approach for dissecting the active ingredients and potential targets of the Chinese Herbal Bufeï Jianpi formula for the treatment of COPD. *Int J Chron Obstruct Pulmon Dis.* 2015;10:2633–2656.
- [16] Zhao P, Li J, Li Y, et al. Integrating transcriptomics, proteomics, and metabolomics profiling with system pharmacology for the delineation of long-term therapeutic mechanisms of Bufeï Jianpi formula in treating COPD. *Biomed Res Int.* 2017;2017:7091087.

- [17] Dong H, Liu X, Zheng W, et al. Three Tiaobu Feishen formulae reduces cigarette smoke-induced inflammation in human airway epithelial cells. *J Tradit Chin Med.* 2020;40(3):386–392.
- [18] Li C, Long Q, Zhang D, et al. Identification of a four-gene panel predicting overall survival for lung adenocarcinoma. *BMC Cancer.* 2020;20(1):1198.
- [19] Shen C, Liu J, Wang L, et al. Identification of metabolism-associated genes and construction of a prognostic signature in bladder cancer. *Cancer Cell Int.* 2020;20(1):538.
- [20] Zhang M, Zhu K, Pu H, et al. An immune-related signature predicts survival in patients with lung adenocarcinoma. *Front Oncol.* 2019;9:1314.
- [21] Zhang L, Chen J, Yang H, et al. Multiple microarray analyses identify key genes associated with the development of non-small cell lung cancer from chronic obstructive pulmonary disease. *J Cancer.* 2021;12(4):996–1010.
- [22] Guan Q, Zeng Q, Yan H, et al. A qualitative transcriptional signature for the early diagnosis of colorectal cancer. *Cancer Sci.* 2019;110(10):3225–3234.
- [23] Guan Q, Zhang J, Guo Y, et al. The effects of age, cigarette smoking, sex, and race on the qualitative characteristics of lung transcriptome. *Biomed Res Int.* 2020;2020:1–7.
- [24] He J, Cheng J, Guan Q, et al. Qualitative transcriptional signature for predicting pathological response of colorectal cancer to FOLFOX therapy. *Cancer Sci.* 2020;111(1):253–265.
- [25] Zhang ZM, Wang JS, Zulfiqar H, et al. Early diagnosis of pancreatic ductal adenocarcinoma by combining relative expression orderings with machine-learning method. *Front Cell Dev Biol.* 2020;8:582864.
- [26] Li J, Zhao P, Yang L, et al. Integrating 3-omics data analyze rat lung tissue of COPD states and medical intervention by delineation of molecular and pathway alterations. *Biosci Rep.* 2017;37(3):BSR20170042.
- [27] Li Y, Li SY, Li JS, et al. A rat model for stable chronic obstructive pulmonary disease induced by cigarette smoke inhalation and repetitive bacterial infection. *Biol Pharm Bull.* 2012;35(10):1752–1760.
- [28] Li Y, Tian YG, Li JS, et al. Bufei Yishen granules combined with acupoint sticking therapy suppress oxidative stress in chronic obstructive pulmonary disease rats: via regulating peroxisome proliferator-activated receptor-gamma signaling. *J Ethnopharmacol.* 2016;193:354–361.
- [29] Mao J, Li Y, Feng S, et al. Bufei Jianpi formula improves mitochondrial function and suppresses mitophagy in skeletal muscle via the adenosine monophosphate-activated protein kinase pathway in chronic obstructive pulmonary disease. *Front Pharmacol.* 2020;11:587176.
- [30] Zhao P, Liu X, Dong H, et al. Bufei Yishen formula restores Th17/Treg balance and attenuates chronic obstructive pulmonary disease via activation of the adenosine 2a receptor. *Front Pharmacol.* 2020;11:1212.
- [31] Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 2000;28(1):27–30.
- [32] Lee KY, Ho SC, Chan YF, et al. Reduced nuclear factor-kappaB repressing factor: a link toward systemic inflammation in COPD. *Eur Respir J.* 2012;40(4):863–873.
- [33] Tanrikulu AC, Abakay A, Evliyaoglu O, et al. Coenzyme Q10, copper, zinc, and lipid peroxidation levels in serum of patients with chronic obstructive pulmonary disease. *Biol Trace Elem Res.* 2011;143(2):659–667.
- [34] Marinari S, Manigrasso MR, De Benedetto F. Effects of nutraceutical diet integration, with coenzyme Q10 (Q-Ter multicomposite) and creatine, on dyspnea, exercise tolerance, and quality of life in COPD patients with chronic respiratory failure. *Multidiscip Respir Med.* 2013;8(1):40.
- [35] Schiffelers SL, Blaak EE, Baarends EM, et al. Beta-adrenoceptor-mediated thermogenesis and lipolysis in patients with chronic obstructive pulmonary disease. *Am J Physiol Endocrinol Metab.* 2001;280(2):357–364.
- [36] Mudunuri U, Che A, Yi M, et al. bioDBnet: the biological database network. *Bioinformatics.* 2009;25(4):555–556.
- [37] Zong D, Ouyang R, Li J, et al. Notch signaling in lung diseases: focus on Notch1 and Notch3. *Ther Adv Respir Dis.* 2016;10(5):468–484.
- [38] Wang Y, Wu Y, Miao X, et al. Silencing of DYRK2 increases cell proliferation but reverses CAM-DR in non-Hodgkin's lymphoma. *Int J Biol Macromol.* 2015;81:809–817.
- [39] Yamashita S, Chujo M, Moroga T, et al. DYRK2 expression may be a predictive marker for chemotherapy in non-small cell lung cancer. *Anticancer Res.* 2009;29(7):2753–2757.
- [40] Meng F, Zhang L, Ren Y, et al. Transcriptome analysis reveals key signature genes involved in the oncogenesis of lung cancer. *Cancer Biomark.* 2020;29(4):475–482.
- [41] Zhang L, Zhang Z, Yu Z. Identification of a novel glycolysis-related gene signature for predicting metastasis and survival in patients with lung adenocarcinoma. *J Transl Med.* 2019;17(1):423.
- [42] Guan Q, Zeng Q, Jiang W, et al. A qualitative transcriptional signature for the risk assessment of precancerous colorectal lesions. *Front Genet.* 2020;11:573787.
- [43] Bodas M, Subramanian B, Moore AR, et al. The NOTCH3 downstream target HEYL is required for efficient human airway basal cell differentiation. *Cells.* 2021;10(11):3215.
- [44] Shi C, Qian J, Ma M, et al. Notch 3 protein, not its gene polymorphism, is associated with the chemotherapy response and prognosis of advanced NSCLC patients. *Cell Physiol Biochem.* 2014;34(3):743–752.
- [45] Ma Y, Li M, Si J, et al. Blockade of Notch3 inhibits the stem-like property and is associated with ALDH1A1 and CD44 via autophagy in non-small lung cancer. *Int J Oncol.* 2016;48(6):2349–2358.
- [46] Kalathiya U, Padariya M, Pawlicka K, et al. Insights into the effects of cancer associated mutations at the UPF2 and ATP-binding sites of NMD master regulator: UPF1. *Int J Mol Sci.* 2019;20(22):5644.
- [47] Rao S, Amorim R, Niu M, et al. The RNA surveillance proteins UPF1, UPF2 and SMG6 affect HIV-1

- reactivation at a post-transcriptional level. *Retrovirology*. 2018;15(1):42.
- [48] Han S, Cao D, Sha J, et al. LncRNA ZFPM2-AS1 promotes lung adenocarcinoma progression by interacting with UPF1 to destabilize ZFPM2. *Mol Oncol*. 2020;14(5):1074–1088.
- [49] Pan Y, Liu H, Wang Y, et al. Associations between genetic variants in mRNA splicing-related genes and risk of lung cancer: a pathway-based analysis from published GWASs. *Sci Rep*. 2017;7:44634.
- [50] Miller CT, Aggarwal S, Lin TK, et al. Amplification and overexpression of the dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 2 (DYRK2) gene in esophageal and lung adenocarcinomas. *Cancer Res*. 2003;63(14):4136–4143.
- [51] Boni J, Rubio-Perez C, Lopez-Bigas N, et al. The DYRK family of kinases in cancer: molecular functions and therapeutic opportunities. *Cancers*. 2020;12(8):2106.
- [52] Shi J, Chen Y, Wang Z, et al. Comprehensive bioinformatics analysis to identify the gene HMMR associated with lung adenocarcinoma prognosis and its mechanism of action in multiple cancers. *Front Oncol*. 2021;11:712795.
- [53] Jiang X, Tang L, Yuan Y, et al. NcrRNA-mediated high expression of HMMR as a prognostic biomarker correlated with cell proliferation and cell migration in lung adenocarcinoma. *Front Oncol*. 2022;12:846536.
- [54] Chen Y, Shen L, Chen B, et al. The predictive prognostic values of CBFA2T3, STX3, DENR, EGLN1, FUT4, and PCDH7 in lung cancer. *Ann Transl Med*. 2021;9(10):843.
- [55] Guan Q, Chen R, Yan H, et al. Differential expression analysis for individual cancer samples based on robust within-sample relative gene expression orderings across multiple profiling platforms. *Oncotarget*. 2016;7(42):68909–68920.
- [56] Zhou YJ, Lu XF, Meng JL, et al. Qualitative transcriptional signature for the pathological diagnosis of pancreatic cancer. *Front Mol Biosci*. 2020;7:569842.
- [57] Cheng J, Guo Y, Gao Q, et al. Circumvent the uncertainty in the applications of transcriptional signatures to tumor tissues sampled from different tumor sites. *Oncotarget*. 2017;8(18):30265–30275.
- [58] Chen R, Guan Q, Cheng J, et al. Robust transcriptional tumor signatures applicable to both formalin-fixed paraffin-embedded and fresh-frozen samples. *Oncotarget*. 2017;8(4):6652–6662.
- [59] Liu H, Li Y, He J, et al. Robust transcriptional signatures for low-input RNA samples based on relative expression orderings. *BMC Genomics*. 2017;18(1):913.