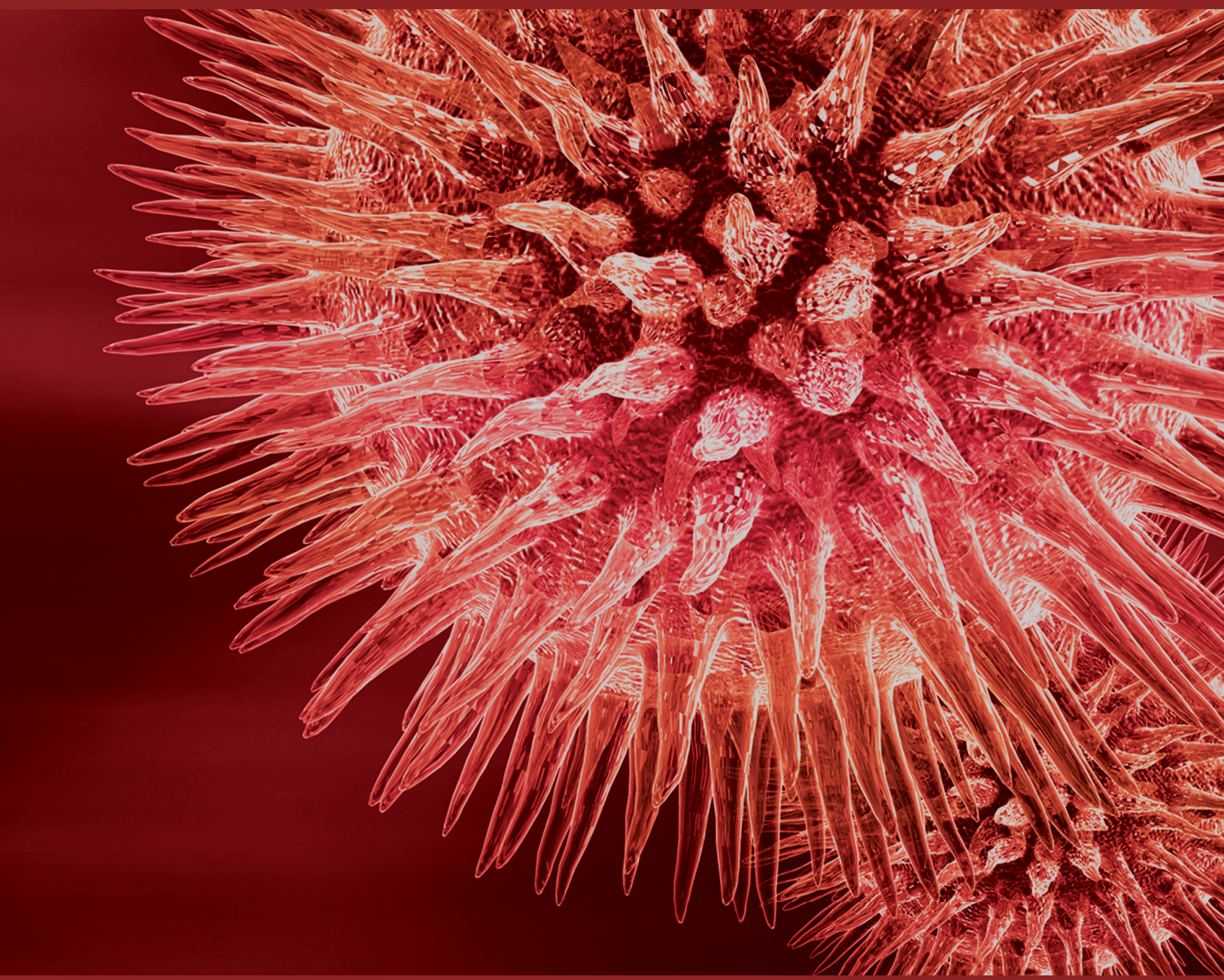
BioMed Research International

Applications of Bioinformatics and Systems Biology in Precision Medicine and Immunooncology

Special Issue Editor in Chief: Yudong Cai
Guest Editors: Tao Huang and Jialiang Yang



Applications of Bioinformatics and Systems Biology in Precision Medicine and Immunooncology

BioMed Research International

Applications of Bioinformatics and Systems Biology in Precision Medicine and Immunooncology

Special Issue Editor in Chief: Yudong Cai
Guest Editors: Tao Huang and Jialiang Yang

Contents

Hindawi
BioMed Research International
Volume 2018, Article ID 1427978, 2 pages
https://doi.org/10.1155/2018/1427978

Editorial

Applications of Bioinformatics and Systems Biology in Precision Medicine and Immunooncology

Yudong Cai,[1] **Tao Huang**,[2] **and Jialiang Yang**[3]

[1]*School of Life Sciences, Shanghai University, Shanghai 200444, China*
[2]*Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China*
[3]*Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York City, NY 10029, USA*

Correspondence should be addressed to Yudong Cai; cai_yud@126.com

Received 25 March 2018; Accepted 25 March 2018; Published 30 September 2018

1. Introduction

Next-Generation Sequencing (NGS) technology, often seen as the foundation of precision medicine, has been successfully applied in oncology diagnostics and immunotherapy. With advances in gene diagnostics and immunotherapy, there may be a chance to control the development of cancers and alleviate the suffering of patients undergoing chemotherapy. To promote the translation of precision medicine from bench to beside and from application of genetic testing to personalized medicine, new analysis methods for NGS and genetic data need to be developed. For example, the NGS panel is quite different from whole genome sequencing (WGS), focusing on fewer genes or regions but requiring greater precision and efficiency. For complex diseases, such as cancers, the driver genes are usually a cluster of genes in a regulatory network. Graph theories, such as shortest path analysis and random walk algorithms, will help dissect genomewide interactions into key modules or paths whose dysfunction is associated with disease progression.

In this special issue, we have received 32 papers, out of which 19 has been accepted for publication. These papers could be generally divided into 3 categories including (1) computational models in identifying key biomarkers, pathways, and network modules associated with cancers and other diseases, (2) validations of the mechanisms of key biomarkers and their applications in tumor diagnosis and treatment, and (3) other studies in predicting tumor evolution, drug-disease association, disease sequence alignment, and so on.

2. Computational Models in Identifying Key Biomarkers

H. Choi and K. J. Na first identified survival-related gene network modules. By selecting representative genes from survival-related modules, they developed a deep learning-based risk stratification model for lung cancer. Their model showed high predictability for prognosis in independent datasets and its predictive value was independent of clinical and pathological features of lung cancer.

J. Zheng et al. proposed a method for constructing the pathway network of gene phenotype. Briefly, it firstly builds a biological pathway network, and then the GeneRank algorithm was used to select disease-associated pathways. The results by gene expression data of breast cancer show that the method proposes an effective way to identify reliable disease-associated pathways.

X. Li et al. presented a new computational method based on the SimRank and density-based clustering recommender model for miRNA-disease association's prediction (SRMDAP). The AUC of 0.8838 based on leave-one-out cross validation and case studies suggested the excellent performance of the SRMDAP in predicting miRNA-disease associations. SRMDAP could also predict diseases without any related miRNAs and miRNAs without any related diseases.

S. Zhu et al. provided a sample-specific method that constructs an index with individual-specific dynamical network biomarkers (DNB), which are defined as early warning index (EWI) for detecting predisease state of individual sample.

Based on microarray data of influenza A disease, 144 genes are selected as DNB and the 7th time period is defined as pre-disease state.

B. Xie et al. first identified differentially expressed genes between patients with poor chemotherapy reaction and patients with good chemotherapy reaction, followed by GO and KEGG pathway analyses, PPI network analysis, survival analysis of hub genes, and miRNA target prediction. This study provides a better understanding about gene expression in drug resistance samples, which is potentially useful for the treatment of osteosarcoma.

3. Validations of Key Biomarkers and Their Applications in Tumor Diagnosis and Treatment

H. Meng et al. investigated the prevalence and prognostic and clinicopathologic features of human papillomavirus-related oropharyngeal cancer in northeast China and elucidated the involvement of p16 in the tumorigenesis and progression of oropharyngeal squamous cell carcinoma (OPSCC) from 1470 OPSCC patients collected from 2000 to 2016. They also demonstrated that p16 expression is significantly associated with early stage primary OPSCCs and the patients with p16 expression tend to show better survival following surgery and radiotherapy.

X. Wu et al. investigated the clinicopathologic and prognostic significance as well as the potential role of *HMGB1* in the development and progression of lung cancer. Their results suggest that *HMGB1* expression is significantly associated with lung cancer progression and might be a potential prognosis and therapeutic marker for lung cancer.

H.-B. Xing et al. investigated whether the knockdown of *IL-6* enhances the gemcitabine sensitivity of PANC-1 cells. In their experiments, the knockdown of *IL-6* induced apoptosis and reduced cell proliferation and tumorigenicity and remarkably promoted the antitumor effect of gemcitabine. The results suggest that combining shRNA targeting *IL-6* and gemcitabine is a potential clinical approach for pancreatic cancer therapy.

R. Liu et al. detected RAS mutation, immunoscore, and PD-L1 expression in 60 Chinese metastatic colorectal cancer (mCRC) patients and suggested PD-L1 expression and RAS status to be prognostic indicators for mCRC patients with palliative operation.

X. Yu et al. investigated the mechanism of AEG-1 for regulating metastasis in hypoxia-induced ovarian carcinoma. By a statistical method comparing the protein amounts of AEG-1, HIF-1α, and VEGF in ovarian cancer tissue specimens, they hypothesized that AEG-1 associates with hypoxia in ovarian cancer by regulating the HIF-1-alpha/NF-kappa-B/ VEGF pathway.

Y. Chen et al. examined PD-L1 and PD-1 expression in thymic epithelial tumor (TET) tissues from patients. They found that PD-L1 expression levels were correlated with disease progression. They also showed that PD-L1 mRNA levels were also correlated with its immunohistochemistry staining levels and thus can be used as alternative method to detect PD-L1 levels in TETs.

X. Wu et al. also revealed that glycyrrhizin reduces the activity of JAK/STAT signaling pathway, which regulates the expression of *HMGB1*. This is a potential mechanism by which glycyrrhizin can inhibit the progression of lung cancer.

Z. Jiang et al. investigated the prognostic value of liver function in colorectal liver metastases patients. They showed that biochemical analyses of liver function tests at the initial diagnosis of colorectal liver metastases enable the stratification of patients into low- and high-risk groups, which may help clinicians to determine promising treatment strategies.

4. Other Studies in Predicting Tumor Evolution, Drug-Disease Association, and Disease Sequence Alignment

Y. Liang et al. proposed an improved binary differential evolution algorithm, BDEP, to infer tumor phylogenetic tree based on fluorescent in situ hybridization (FISH) platform. The constructed phylogenetic trees have great performance in characterizing tumor development process, which outperforms other similar algorithms.

H. Cui et al. developed a method called PEDD to predict the effect of drugs on diseases by using gene expression profiles from both disease-related tissues and cell lines treated with drugs. The authors also implemented the method with an interactive web tool and applied the method to real microarray and RNA-seq datasets.

Finally, M. Li et al. proposed a novel protein sequence alignment method, which outperforms a few current alignment methods in accuracy. Y. Luo et al. developed a computational method to investigate key GO terms and KEGG pathways associated with copy number variations (CNVs). Y. Fang et al. explored the genetic polymorphism of *Aedes albopictus* and its association with diseases like dengue. And L. Wang et al. studied the roles of swine as a mixtures for influenza viruses.

In summary, we expect that this special issue updates novel NGS-based computational methods in predicting key genes, pathways, and network modules associated with diseases like cancers and experimental validations of their mechanisms, as well as their applications in disease diagnosis and treatment. The studies will serve as a bridge to connect computational models in mining clinical NGS data and translation of the findings into personalized therapies for diseases.

Acknowledgments

We are grateful to the authors for contributing their valuable work to this special issue and the reviewers for this constructive comments. We also thank the editorial board for approving this topic and hope this issue will advance the research in disease-associated biomarker identification and its applications in translational research.

Yudong Cai
Tao Huang
Jialiang Yang

Hindawi
BioMed Research International
Volume 2018, Article ID 9594568, 8 pages
https://doi.org/10.1155/2018/9594568

Research Article

Association of p16 as Prognostic Factors for Oropharyngeal Cancer: Evaluation of p16 in 1470 Patients for a 16 Year Study in Northeast China

Hong-xue Meng,[1] Su-sheng Miao,[2] Kexin Chen,[1] Hui-ning Li,[3,4] Guodong Yao,[1] Jiashi Geng,[5] Hongmei Wang,[1] Qing-tao Shi,[1] Jing He,[1] Xionghui Mao,[2] Fang-jia Tong,[6] Lan-Lan Wei,[6] Ji Sun,[2] Dongfeng Tan,[1] Qi You,[7] Xiaomei Li ,[1] and Jing-shu Geng [1]

[1]*Department of Pathology, Harbin Medical University Cancer Hospital, Harbin, China*
[2]*Department of Otolaryngology, Head and Neck Surgery, Harbin Medical University Cancer Hospital, Harbin, China*
[3]*Department of Pathology, The First Affiliated Hospital of Heilongjiang University of Chinese Medicine, Harbin, China*
[4]*Department of pathology, Harbin Medical University, Harbin, China*
[5]*Department of Radiology, Harbin Medical University Cancer Hospital, Harbin, China*
[6]*Department of Microbiology, Harbin Medical University, Harbin, China*
[7]*Department of Gastroenterology, Harbin Medical University Cancer Hospital, Harbin, China*

Correspondence should be addressed to Xiaomei Li; fanliwenqi@163.com and Jing-shu Geng; gengjingshu@yeah.net

Received 28 November 2017; Revised 22 December 2017; Accepted 23 January 2018; Published 17 September 2018

Academic Editor: Tao Huang

Human papillomavirus (HPV) is an etiological risk factor for oropharyngeal squamous cell carcinomas (OPSCC). Our study investigates the prevalence, prognostic, and clinicopathologic features of HPV-related oropharyngeal cancer in Northeast China and elucidates the involvement of p16 in the tumorigenesis and progression of OPSCC. Specimens from 1470 OPSCC patients collected from 2000 to 2016 were analyzed using the status of HPV by polymerase chain reaction (PCR) and p16 immunohistochemistry. Overexpression of p16 was observed in 81 (5.51%) of the 1470 cases, and HPV positive was present in 78 cases (5.31%) of the 1470 cases. HPV positive and p16 overexpression have a good concordance. However, we found that the etiological fraction of HPV in cancers of the OPSCCs was obviously lower in Northeast China than other cohorts previously reported. Interestingly, nearly 89% of patients with p16 expression were smokers, and nearly 70% of patients with p16 expression had a history of alcohol. Our study also demonstrates that p16 expression is significantly associated with early stage primary OPSCCs and the patients with p16 expression tend to show better survival following surgery and radiotherapy.

1. Introduction

Head and neck squamous cell carcinoma (HNSCC) has been defined as the sixth leading cause of cancer in the world [1]. High recurrence rates and nodal metastases always lead to high mortality of HNSCC. Especially, 5-year survival rates of HNSCC patients with cervical lymph node metastases are reduced by approximately 50% [2]. Conventionally, patients diagnosed with early stage HNSCC would have good prognosis after surgery and adjuvant radiation [3, 4].

Before HPV positive as a new risk factor for HNSCC was found, many risk factors had been reported, including tobacco, poor oral hygiene, and alcohol [5, 6]. Then the prevalence of HPV-related HNSCC, especially oropharyngeal squamous cell carcinomas (OPSCC), was largely observed in many populations in Western Europe, United States, and Australia [3–7]. Nonetheless, the prevalence, prognostic, significance, and correlations of high-risk HPV infection in OPSCCs in China cohort, accounting for 1/4 of the global population, remain blurry. And the precise pathogenesis and clinic pathologic features of HPV-related oropharyngeal cancer in Northeast China are still unclear.

High-risk human papillomavirus (HPV) infection causes the increase of OPSCC [8]. Many studies have shown that

the prevalence of HPV-related OPSCC has been evaluated to range from 45 to 90% [3–7]. Moreover, a dominant subtype of HPV16 is thought to represent 90% of HPV-related OPSCC. HPV-related OPSCC is identified as a unique clinical entity. Patients with HPV associated SCC are expected to have the improved survival. Thus the clinical value of exploring the role of HPV in OPSCC is also beneficial to decrease treatment related side-effect [8].

p16 protein expression has been reported to be related to HPV infection, and p16 may be used as a predictive biomarker for HPV high-risk tumors [9]. p16 as a cyclin-dependent kinase inhibitor played an important role in inhibiting CDK4 and cyclin D1 complex dependent phosphorylation of Rb (retinoblastoma), as a tumor suppressor protein [10]. Viral oncoproteins E7 is always expressed in HPV-related cancers. Studies had shown an inhibitory effect of E7 on Rb activation by HPV infection [11, 12]. And inactivation of Rb by HPV-expressed E7 induced the transcription of the cyclin-dependent kinase inhibitor p16 [13]. Importantly, the expression of p16 was a positive indicator for improved survival. Several researches have demonstrated that p16 was a more effective independent prognostic factor for overall survival and progression-free survival than HPV status prediction [14, 15]. However, whether p16 immunohistochemistry could be used as a strong discriminator of clinical outcome in patients with OPSCC has not been defined. Larger studies are necessary to determine whether p16 can be used as well established prognostic variables, including T category, depth of invasion, and nodal status of OPSCC.

In our study, we first investigated the prevalence and prognostic and clinicopathologic features of HPV-related oropharyngeal cancer in Northeast China. Furthermore, we observed that p16 expression was significantly associated with early stage primary OPSCCs and that patients with p16 expression tend to show better survival following surgery and radiotherapy. Our results suggest that p16 may be a prognostic factor of OPSCCs in China.

2. Methods

2.1. Patients. This study enrolled 1470 patients with pathology-proven oropharyngeal cancer. Patients were recruited from Harbin Medical University Cancer Hospital (Cancer Center for Northeast China, Harbin, China) from January 2000 to February 2016. Tissues were obtained from patients during surgery.

2.2. Ethics Statement. According to the principles of the Declaration of Helsinki, we conducted this research. All participants in this study signed the written informed consent. The study had been approved by the Institutional Ethics Committee of Harbin Medical University Cancer Hospital.

2.3. Clinical Parameters. The clinical data of controls and IgAN patients, including age, history of smoking, gender, history of alcohol, and treatment, were collected.

2.4. Histopathological Diagnosis. All cases were diagnosed and categorized according to the WHO classification. All slides were reviewed by two pathologists and scored the pathological variables. International Collaboration on Oropharyngeal Cancer Network for Staging (ICON-S) has developed a TNM classification specific to HPV positive oropharyngeal cancer [16, 17]. We followed the TNM stage from 7th edition of the UICC/AJCC TNM classification: no lymph nodes as ICON-S N0; ipsilateral lymph nodes as ICON-S N1; bilateral or contralateral lymph nodes as ICON-S N2; lymph nodes larger than 6 cm as ICON-S N3, which resembles the N classification of nasopharyngeal carcinoma except the lack of a lower neck lymph node variable. The proposed ICON-S classification is as follows: stage I is T1-T2N0-N1, stage II is T1-T2N2 or T3N0–N2, and stage III is T4 or N3. Metastatic disease (M1) is classified as ICON-S stage IV.

2.5. Antibodies and Immunohistochemistry (IHC). Formalin-fixed samples and paraffin-embedded sections (4 μm thick) were first blocked with 1% H_2O_2. Then the samples were treated by antigen retrieval in trypsin for 30 min at 37°C, followed by immersion in citrate buffer (pH 6.0; Mitsubishi Chemical Medience, Tokyo, Japan) for 20 min at 120°C in an autoclave. Protein Blocking Agent (Streptavidin-Biotin Universal Detection System; Beckman Coulter, Marseille, France) was used to block the sections. And then the sections were incubated with the following primary antibodies overnight at 4°C: rabbit anti-human P16 (1 : 100, INK4a, IgG, Zhongshan, China). After that, sections were incubated with secondary antibodies from the Streptavidin-Biotin Universal Detection System (Beckman Coulter) and visualized by DAB. The negative controls were specific isotype control antibodies and phosphate-buffered saline (PBS; omitting primary antibodies).

For calculating the p16 [INK4a] expression, nuclear and cytoplasmic positivity were identified as positive reactions and were scored semiquantitatively as described by previous study [15]: negative score was <1% of positive cells; sporadic score was that isolated cells were positive but <5%; focal score was small cell clusters but <80% of positive cells; and diffuse score was >80% of positive cells. Positive cells with p16 expression were defined as strong and diffuse nuclear and cytoplasmic staining in at least 80 percent or more of the tumor cells.

2.6. DNA Extraction and PCR Analysis. Total DNA was extracted and purified from formalin-fixed, paraffin-embedded tissues by DNeasy Micro kit (Qiagen, Hilden, Germany). The resulting DNA was amplified for 35 cycles by PCR. The forward and reverse primers were listed as follows: β-globin 5′-GAA GAG CCA AGG ACA GGT AC-3′ (forward) and 5′-CAA CTT CAT CCA CGT TCA CC-3′ (reverse); HPV 5′-CGT CCM ARR GGA WAC TGA TC-3′ (forward) and 5′-GCM CAG GGW CAT AAY AAT GG-3′ (reverse).

2.7. Statistical Analysis. Student's t-test was performed to estimate the significant difference between HPV positive OPSCC patients and HPV negative OPSCC patients. We also analyzed the correlation among clinical presentation, HPV state, and P16 by Spearman's correlation analysis and

Table 1: Profiles and clinical parameters of patients.

Clinicopathological findings		
Variable	*n*	%
Age at diagnosis, years		
⩽45	85	5.78
46–55	469	31.9
56–65	624	42.45
⩾66	292	19.86
Mean (SD)	58.24 ± 6.64	
Age range	31–86	
Sex		
Male	1167	79.39
Female	303	20.61
History of smoking		
Yes (current/former)	1296	88.16
No (never)	174	11.84
History of alcohol		
Yes (current/former)	890	60.54
No (never)	580	39.46
Treatment		
Surgery alone	1257	85.51
Surgery + radiotherapy	205	13.95
Surgery + chemoradiotherapy	5	0.34
Surgery + radiotherapy + chemoradiotherapy	3	0.2
Event after initial CRT		
Residual tumor (PD, SD, and PR)	286	19.46
CR followed by recurrence/metastasis	441	30
Durable CR	743	50.54

Continuous variables are given as mean ± standard deviation.

Pearson's correlation analysis (SAS Institute Inc., Cary, NC, USA). *P* values less than 0.05 were considered as significant differences.

3. Results

3.1. Clinical and Pathological Parameters. The clinicopathological characters of the 1470 cases of OPSCC were represented in Table 1. Most of patients were males ($n = 1167$; 79.39%) and smokers ($n = 1296$; 88.16%). About sixty percent (60.54%) of 890 patients were alcohol consumers. Among all these patients, patients with surgery only were 1257 cases, patients with surgery followed by radiotherapy were 205 cases, patients with surgery followed by chemotherapy were 5 cases, and patients with surgery followed by radiotherapy and chemotherapy were 3 cases. All of the patients had available follow-up information. 286 patients (19.46%) presented with residual disease, but 1184 patients (80.54%) initially obtained a complete response (CR) after finishing the initial CRT. 62.75% (743/1184) of patients maintained the CR during follow-up; however, 37.24% (441/1184) of patients subsequently showed recurrence or metastasis.

Table 2: Histological diagnosis of patients.

Histological diagnosis	*n*	%
Squamous cell carcinoma		
SCC NOS/conventional nonkeratinizing	174	11.84
Conventional keratinizing	1271	86.46
Conventional exophytic keratinizing	14	0.95
Basaloid/papillary	2	0.14
Verrucous	1	0.07
Sarcomatoid	4	0.27
Undifferentiated carcinoma	2	0.14
Adenosquamous carcinoma	2	0.14
Differentiation		
Well	467	31.77
Moderate	818	55.65
Poor	185	12.59
Lymphovascular invasion	615	41.83
Perineural invasion	441	30
Extracapsular spread	244	16.6
Bone invasion	148	10.07
Pathological T category		
T1	182	12.38
T2	1213	82.52
T3	61	4.15
T4	14	0.95
Pathological N category		
N0	855	58.16
N1	422	28.71
N2a	119	8.1
N2b	57	3.88
N2c	17	1.16
N3	0	
Clinical stage		
I/II	1338	91.02
III/IV	132	8.98

According to the histologic typing, 174 (11.84%) of the 1470 cases were SCC NOS/conventional nonkeratinizing and 1271 (86.46%) cases were conventional keratinizing. For the differentiation, 467 (31.77%) cases were well, 818 (55.65%) cases were moderate, and 185 (12.95%) cases were poor. Most cases (615, 41.83%) have lymphovascular invasion. For TNM stage statistic, 1395 patients had low–T stage (T1/T2) OPSCC tumors and 75 patients had high–T stage (T3/T4) OPSCC tumors. Moreover, patients with clinically positive lymph node metastasis (N+) were 615 (41.84%). For clinical stage statistic, patients with low clinical stage (I/II) were 91.02% (1338/1470) and high clinical stage (III/IV) were 8.98% (132/1470) (Table 2).

3.2. The Relationship between p16 Protein Overexpression and HPV Status. 5.51% (81/1470) of OPSCC samples were detected to p16 overexpression by immunohistochemistry (Figure 1). HPV was positive in 78 cases (5.31%) of the 1470 cases by PCR (Figure 2).

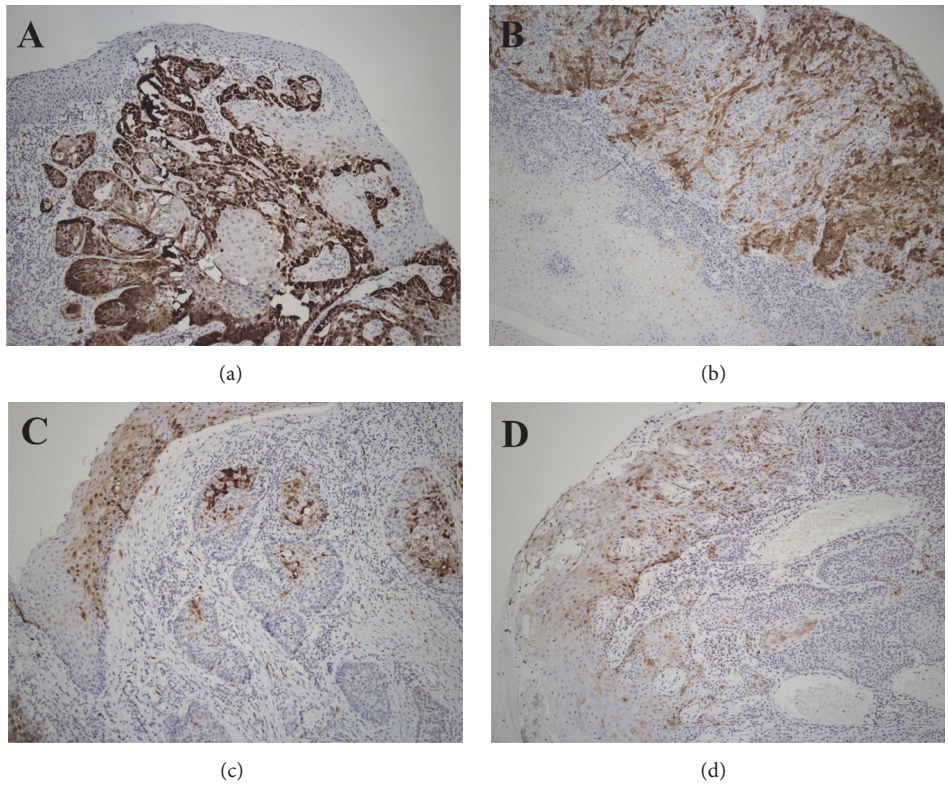


Figure 1: *Expression of P16-positive cells in oropharyngeal cancer*. Immunohistochemical analysis was used to show the expression of P16-positive cells in oropharyngeal cancer (magnification: ×400); nuclear and cytoplasmic positivity were classified as positive reactions and were scored as (a) diffuse (>80% of the cells were stained); (b) focal (small cell clusters, but <80% of the cells were positive); (c) sporadic (isolated cells were positive but <5%); (d) negative (<1% of cells were positive).

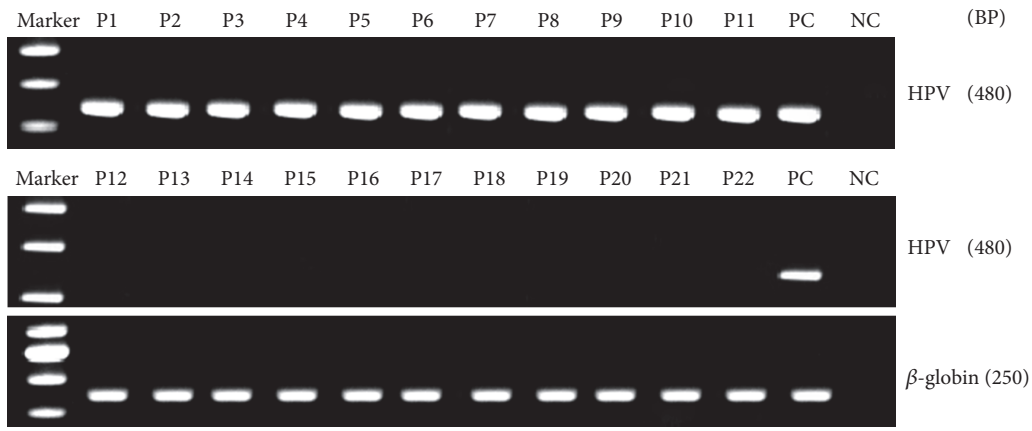


Figure 2: *HPV DNA-PCR of oropharyngeal cancer.* Using PCR, HPV state was detected in Oropharyngeal cancer patients.

Good concordance between HPV positive status and p16 overexpression was established, which was with high sensitivity (100%) and high specificity (96%; Table 3). Consistently, we found that HPV status was significantly more frequently present among the young (age: 46–55) ($P < 0.01$), males ($P < 0.05$), conventional keratinizing type ($P < 0.01$), moderate and poor differentiation ($P < 0.05$), low T stage ($P < 0.05$), lymph node metastasis ($P < 0.05$), high clinical stage ($P < 0.05$), and p16 overexpression ($P < 0.01$) cases (Tables 3 and 4). Specifically, HPV and p16 positive patients usually maintained the CR more during follow-up. Our results also indicated that p16 expression may be a prognostic marker with an improved response to both radiation therapy and chemotherapy.

3.3. p16 Expression Was Significantly Associated with Improved Survival. Among all 1470 cases, positive p16 expression was linked with markedly improved overall survival (OS,

Table 3: The relationship between p16 overexpression, HPV status, and clinical parameters of patients.

Variable	p16-IHC $n = 1470$		P	HPV DNA-PCR $n = 1470$		P
	Positive ($n = 81$)	Negative ($n = 1389$)		Positive ($n = 78$)	Negative ($n = 1392$)	
Patient characteristics						
Age at diagnosis, years						
⩽45	12	73	0.0034	11	74	0.0012
46–55	43	426	0.00004	41	428	0.000058
56–65	16	608	0.00002	16	608	0.000056
⩾66	10	282	0.081	10	282	0.109
Sex						
Male	72	1095	0.029	69	1098	0.0417
Female	9	294	0.041	9	294	0.041
History of smoking						
Yes (current/former)	72	1224	0.835	69	1227	0.933
No (never)	9	165	0.833	9	165	0.833
History of alcohol						
Yes (current/former)	57	833	0.062	54	836	0.106
No (never)	24	556	0.082	24	556	0.08
Treatment						
Surgery alone	57	1200	0.00006	55	1202	0.00011
Surgery + radiotherapy	24	181	0.00003	23	182	0.000047
Surgery + chemoradiotherapy	0	8	0.49	0	8	0.501
Event after initial CRT						
Residual tumor (PD, SD, and PR)	21	265	0.13	21	265	0.086
CR followed by recurrence/metastasis	35	406	0.007	33	408	0.014
Durable CR	25	718	0.0003	24	719	0.00033

Table 4: The relationship between p16 overexpression, HPV status, and Histological diagnosis.

Variable	p16-IHC $n = 1470$		P	HPV DNA-PCR $n = 1470$		P
	Positive ($n = 81$)	Negative ($n = 1389$)		Positive ($n = 78$)	Negative ($n = 1392$)	
Histological diagnosis						
Squamous cell carcinoma						
SCC NOS/conventional nonkeratinizing	13	161	0.862	12	162	0.318
Conventional keratinizing	66	1205	0	64	1207	0.242
Conventional exophytic keratinizing	2	12	0.169	2	6	0.127
Basaloid/papillary	0	2	0.732	0	2	0.737
Verrucous	0	1	0.809	0	1	0.813
Sarcomatoid	0	4	0.628	0	4	0.635
Undifferentiated carcinoma	0	2	0.732	0	2	0.737
Adenosquamous carcinoma	0	2	0.732	0	2	0.737
Differentiation						
Well	24	443	0.0017	23	444	0.656
Moderate	20	798	0	19	799	0
Poor	37	148	0.000002	36	149	0
Lymphovascular invasion	56	559	0.0027	56	559	0
Perineural invasion	77	364	0.036	76	365	0
Extracapsular spread	10	234	0.068	8	236	0.122
Bone invasion	4	144	0	4	144	0.136
Pathological T category						
T1	4	178	0.073	4	178	0.0456
T2	73	1140	0	70	1143	0.084
T3	2	59	0.394	2	59	0.471
T4	1	13	0	1	13	0.758
Pathological N category						
N0	26	829	0.806	24	831	0
N1	25	397	0.02	24	398	0.679
N2a	19	100	0.00066	19	100	0
N2b	9	48	0.00052	9	48	0.00031
N2c	2	15	0.255	2	15	0.232
N3	0	0		0	0	
Clinical stage						
I/II	6	126	0.611	6	126	0.683
III/IV	75	1263	0.611	72	1266	0.683

Table 5: Association between p16 status and survival.

	Univariable (p value)	Multivariable, HR (95% CI) P value	In patients receiving radiotherapy (P value)
Disease-specific	0.03[a]	0.19 (0.02–1.38) 0.100[b]	0.14[a]
Disease-free	0.14[a]	0.67 (0.32–1.36) 0.266	0.014[a]
Overall survival	0.05[a]	0.44 (0.15–1.23) 0.118	0.038[a]

[a]Log rank test P value; [b]adjusting for effect of depth of invasion alone.

$P = 0.05$), but this result was not significant in multivariate analysis (Table 5).

4. Discussion

This study including 1470 patients over a period of 16 years has a few inherent biases typical of a retrospective cohort. The principles of surgical management and patient selection for radiotherapy have essentially remained the same during this period, although there has been increasing use of adjuvant chemotherapy and highly conformal radiation techniques. Consequently, we have used HPV status detection as the gold standard to evaluate the potential clinical value of other prognostic markers for HPV-related OPSCC.

The percentage of HPV positive OPSCC in northeast Chinese patients calculated by our research was substantially lower than that published by recent meta-analysis data in oropharyngeal cancers [18]. The percentage of HPV positive OPSCC was 11.7% in an Eastern Chinese population and 21.7% in Southern Chinese patients as previously reported [19, 20]. The discrepancy may be related to the differences in the geographic origin of patients, heterogeneous laboratory procedures, or different methods used to detect HPV status.

We also found that HPV status was significantly correlated to sex and age of patients. About sex, HPV positive estimates were substantially higher in men than in women in our cohort, different from European cohort in other studies [3–7]. Additionally, we also found that HPV positive estimates were substantially higher in 46–55 ages. Finally, the incidence of tobacco smoking was 88.16% ($n = 1296$) in our cohort and nearly 89% of patients with p16 expression were smokers. Sixty (60.54%) of 890 patients were alcohol consumers, and nearly 70% of patients with p16 expression had a history of alcohol. The carcinogenic effects of smoking and alcohol mediated through p53 mutations are notable.

In our results, we observed that OPSCC patients with p16 overexpression had significantly longer disease-specific survival than p16 negative patients following surgery as well as postoperative adjuvant radiotherapy, which was consistence with published data about the potential prognostic maker of p16 in oropharyngeal cancer. It suggested that p16 may be a biomarker for predicting the prognosis of OPSCC patients in China.

HPV and p16 as biomarkers or therapeutic targets in the treatment of HNSCC have the growing consensus of the importance [21]. In our study, p16 positive patients had significantly longer disease-specific survival on univariable analysis, which was essentially equivalent to that published by previous reports [22–24]. However, p16 expression was not only an independent predictor of survival on multivariable analysis. As discussed above, it may be reasonable to assume that p16 expression may also mediate survival of OPSCC patients by controlling the proliferative capacity and invasive potential of the primary tumor.

5. Conclusion

Our study demonstrates that p16 expression is significantly associated with early stage primary OPSCCs and that patients with p16 expression tend to show better survival following surgery and radiotherapy. p16 expression, as well as HPV status, may be a prognostic maker of OPSCCs in China. Furthermore, the etiological fraction of HPV in cancers of the OPSCCs is substantially lower in Northeast China than that in United States and Western Europe. Thus, the real prevalence of HPV in OPSCCs is still the future burden. Further researches will define the more detailed mechanisms underlying HPV involvement in OPSCCs.

Conflicts of Interest

All authors have read the journal's policy on disclosure of potential conflicts of interest declare that they have no conflicts of interest.

Authors' Contributions

Hong-xue Meng, Su-sheng Miao, and Kexin Chen contributed equally to this study.

Acknowledgments

This work was supported National Nature Science Foundation of China (81600539, 81372785, 81400443, and 81372178), Natural Science Foundation of Heilongjiang Province of China (QC2012C041, LC2016038), Foundation of Heilongjiang administration of Traditional Chinese Medicine (Huining Li, ZHY16-032), Chinese Postdoctoral Science Foundation (2015M581472), Special Financial Grant from the China Postdoctoral Science Foundation (2016T90310), Postdoctoral Science Foundation of Heilongjiang Province of China (LBH-Z16101, LBH-TZ0616), Heilongjiang Human

Resources and Social Security Bureau (Hong-xue Meng), Harbin Special Fund Project for Science and Technology Innovation (2016RAQXJ203), Foundation for Liver and Gall Group of Nn10 Fund Project of Harbin Medical University Cancer Hospital, and Youth Elite Training Foundation of Harbin Medical University Cancer Hospital (JY2016-06).

References

[1] W. L. Wong, J. Dunn, and H. Mehanna, "The role of PET CT in the management of advanced nodal head neck cancer post chemoradiotherapy," *Translational Cancer Research*, vol. 5, pp. S932–S932, 2016.

[2] S. Z. Liu, D. P. Zandberg, L. M. Schumaker, J. C. Papadimitriou, and K. J. Cullen, "Correlation of p16 expression and HPV type with survival in oropharyngeal squamous cell cancer," *Oral Oncology*, vol. 51, no. 9, article no. 3263, pp. 862–869, 2015.

[3] Cancer Genome Atlas Network, "Comprehensive genomic characterization of head and neck squamous cell carcinomas," *Nature*, vol. 517, no. 7536, pp. 576–582, 2015.

[4] A. Panwar, R. Batra, W. M. Lydiatt, and A. K. Ganti, "Human papilloma virus positive oropharyngeal squamous cell carcinoma: a growing epidemic," *Cancer Treatment Reviews*, vol. 40, no. 2, pp. 215–219, 2014.

[5] H. J. Ryu, E. K. Kim, S. J. Heo, B. C. Cho, H. R. Kim, and S. O. Yoon, "Architectural patterns of p16 immunohistochemical expression associated with cancer immunity and prognosis of head and neck squamous cell carcinoma," *APMIS-Acta Pathologica, Microbiologica et Immunologica Scandinavica*, vol. 125, no. 11, pp. 974–984, 2017.

[6] E. Gelwan, I. Malm, A. Kharanjian, C. Fakhry, J. A. Bishop, and W. H. Westra, "Nonuniform Distribution of High-risk Human Papillomavirus in Squamous Cell Carcinomas of the Oropharynx," *The American Journal of Surgical Pathology*, vol. 41, no. 12, pp. 1722–1728, 2017.

[7] A. J. Hobbs, N. T. Brockton, T. W. Matthews et al., "Primary treatment for oropharyngeal squamous cell carcinoma in Alberta, Canada: A population-based study," *Head & Neck*, vol. 39, no. 11, pp. 2187–2199, 2017.

[8] J. Mallen-St Clair, M. Alani, M. B. Wang, and E. S. Srivastan, "Human papillomavirus in oropharyngeal cancer: The changing face of a disease," *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, vol. 1866, no. 2, pp. 141–150, 2016.

[9] P. R. Aguiar Pastrez, V. S. Mariano, A. M. Costa et al., "The relation of HPV infection and expression of p53 and p16 proteins in esophageal squamous cells carcinoma," *Journal of Cancer*, vol. 8, no. 6, pp. 1062–1070, 2017.

[10] H. H. Al-Khalaf, S. C. Nallar, D. V. Kalvakolanu, and A. Aboussekhra, "p16INK4A enhances the transcriptional and the apoptotic functions of p53 through DNA-dependent interaction," *Molecular Carcinogenesis*, vol. 56, no. 7, pp. 1687–1702, 2017.

[11] A. M. Mills, D. C. Dirks, M. D. Poulter, S. E. Mills, and M. H. Stoler, "HR-HPV E6/E7 mRNA in situ hybridization. Validation against PCR, DNA in situ hybridization, and p16 immunohistochemistry in 102 samples of cervical, vulvar, anal, and head and neck neoplasia," *The American Journal of Surgical Pathology*, vol. 41, no. 5, pp. 607–615, 2017.

[12] K. A. Lang Kuhs, A. R. Kreimer, S. Trivedi et al., "Human papillomavirus 16 E6 antibodies are sensitive for human papillomavirus-driven oropharyngeal cancer and are associated with recurrence," *Cancer*, vol. 123, no. 22, pp. 4382–4390, 2017.

[13] T. Nakagawa, K. Matsusaka, K. Misawa et al., "Frequent promoter hypermethylation associated with human papillomavirus infection in pharyngeal cancer," *Cancer Letters*, vol. 407, pp. 21–31, 2017.

[14] E.-S. Prigge, C. Toth, G. Dyckhoff et al., "P16INK4a/Ki-67 coexpression specifically identifies transformed cells in the head and neck region," *International Journal of Cancer*, vol. 136, no. 7, pp. 1589–1599, 2015.

[15] K. K. Ang, J. Harris, R. Wheeler, R. Weber et al., "Human Papillomavirus and Survival of Patients with Oropharyngeal Cancer," *The New England Journal of Medicine*, vol. 363, no. 1, pp. 24–35, 2010.

[16] S. H. Huang, W. Xu, J. Waldron et al., "Refining American joint committee on cancer/union for international cancer control TNM stage and prognostic groups for human papillomavirus-related oropharyngeal carcinomas," *Journal of Clinical Oncology*, vol. 33, no. 8, pp. 836–845, 2015.

[17] K. R. Dahlstrom, A. S. Garden, W. J. N. William, M. Y. Lim, and E. M. Sturgis, "Proposed staging system for patients with hpv-related oropharyngeal cancer based on nasopharyngeal cancer n categories," *Journal of Clinical Oncology*, vol. 34, no. 16, pp. 1848–1854, 2016.

[18] A. K. Chaturvedi, W. F. Anderson, J. Lortet-Tieulent et al., "Worldwide trends in incidence rates for oral cavity and oropharyngeal cancers," *Journal of Clinical Oncology*, vol. 31, no. 36, pp. 4550–4559, 2013.

[19] Z. Wang, R.-H. Xia, D.-X. Ye, and J. Li, "Human papillomavirus 16 infection and TP53 mutation: Two distinct pathogeneses for oropharyngeal squamous cell carcinoma in an eastern Chinese population," *PLoS ONE*, vol. 11, no. 10, Article ID e0164491, 2016.

[20] E. W. H. Lam, J. Y. W. Chan, A. B. W. Chan et al., "Prevalence, clinicopathological characteristics, and outcome of human papillomavirus-associated oropharyngeal cancer in southern Chinese patients," *Cancer Epidemiology, Biomarkers & Prevention*, vol. 25, no. 1, pp. 165–173, 2016.

[21] D. Rischin, R. J. Young, R. Fisher et al., "Prognostic significance of p16INK4Aand human papillomavirus in patients with oropharyngeal cancer treated on TROG 02.02 phase III trial," *Journal of Clinical Oncology*, vol. 28, no. 27, pp. 4142–4148, 2010.

[22] A. Al-Kaabi, L. W. van Bockel, A. J. Pothen, and S. M. Willems, "$P16^{INK4A}$ and $p14^{ARF}$ gene promoter hypermethylation as prognostic biomarker in oral and oropharyngeal squamous cell carcinoma: a review," *Disease Markers*, vol. 2014, Article ID 260549, 8 pages, 2014.

[23] P. Molony, N. Kharytaniuk, S. Boyle et al., "Impact of positive margins on outcomes of oropharyngeal squamous cell carcinoma according to p16 status," *Head & Neck*, vol. 39, no. 8, pp. 1680–1688, 2017.

[24] F. Wang, H. Zhang, Y. Xue et al., "A systematic investigation of the association between HPV and the clinicopathological parameters and prognosis of oral and oropharyngeal squamous cell carcinomas," *Cancer Medicine*, vol. 6, no. 5, pp. 910–917, 2017.

Hindawi
BioMed Research International
Volume 2018, Article ID 5670210, 12 pages
https://doi.org/10.1155/2018/5670210

Research Article

Pathway Network Analysis of Complex Diseases Based on Multiple Biological Networks

Fang Zheng, Le Wei, Liang Zhao, and FuChuan Ni

College of Informatics, Huazhong Agricultural University, Wuhan 430079, China

Correspondence should be addressed to FuChuan Ni; fcni_cn@mail.hzau.edu.cn

Received 10 November 2017; Revised 6 February 2018; Accepted 11 March 2018; Published 30 July 2018

Academic Editor: Tao Huang

Biological pathways play important roles in the development of complex diseases, such as cancers, which are multifactorial complex diseases that are usually caused by multiple disorders gene mutations or pathway. It has become one of the most important issues to analyze pathways combining multiple types of high-throughput data, such as genomics and proteomics, to understand the mechanisms of complex diseases. In this paper, we propose a method for constructing the pathway network of gene phenotype and find out disease pathogenesis pathways through the analysis of the constructed network. The specific process of constructing the network includes, firstly, similarity calculation between genes expressing data combined with phenotypic mutual information and GO ontology information, secondly, calculating the correlation between pathways based on the similarity between differential genes and constructing the pathway network, and, finally, mining critical pathways to identify diseases. Experimental results on Breast Cancer Dataset using this method show that our method is better. In addition, testing on an alternative dataset proved that the key pathways we found were more accurate and reliable as biological markers of disease. These results show that our proposed method is effective.

1. Introduction

The growth in knowledge of large-scale transcriptomic and proteomic technologies has enabled the identification of risk factors of complex diseases, personalized medicine, and so forth. Many algorithms (such as the supervision, nonsupervision, and statistics method) have been developed to process these data for acquiring important biological biomarkers. However, these methods still have some limitations and challenges. First, transcriptomic data analysis is the inherent complexity of multiple biological processes. Second, the data from different platforms also lead to noise. Although some methods were used to reduce the deviation, it is difficult to obtain robust result of these data. To circumvent these limitations, some computational methods project gene expression data into a molecular signaling network, but errors generated by variations in experiment also affect the accuracy to distinct different samples.

Biological networks are powerful resources for the discovery of genes and genetic modules that drive disease. Fundamental to network analysis is the concept that genes underlying the same phenotype tend to interact; this principle can be used to combine and to amplify signals from individual genes [1], a biological pathway which plays an important role in understanding the mechanisms of complex diseases, improving clinical treatment, and discovering drug targets and biomarkers [2]. The increasing availability of high-throughput biological data of complex diseases and the development of various biological networks provided better conditions to build accurate pathway analysis models. But due to a lack of abundant pathway knowledge, most pathway analysis results are incomplete, unreliable, or inaccurate [3]. So, it is an important work to build pathway analysis method of multitype data, such as gene expression data, transcriptomic data, and protein data. The major advantage of pathway-based methods is their capability to perform biologically relevant dimension reduction as a result of the analysis [4].

There are some popular pathway analysis tools, such as GSEA, SPIA, DAVID, and Pathologist. They provide different methods to explain the function of the pathway analysis.

GSEA (Gene Set Enrichment Analysis) [5] processes expression profile data with labels, sorting every pathway according to enrichment statistics of the difference of gene.

SPIA (Significant Pathway Inference Analysis) [6] combines the difference expression and pathway structure information. The effect of alteration of gene expression at different positions in the pathway is considered to be different.

DAVID [7] (Database for Annotation, Visualization and Integrated Discovery) can provide a comprehensive functional annotation of the gene list to help researchers understand the biological significance of genes behind. At present, it is the most widely used method of gene function annotation.

Although many of the above methods have shown encouraging results for finding new information, there are still some limitations. In those methods, pathway is simplified into a simple set of genes, treating pathways as unstructured sets of genes, ignoring the functional relationship between different genes in the pathway, so it cannot accurately assess function change of the related pathways.

To overcome the above limitations, new analytical methodologies are required that infer complex transcriptomic changes more accurately into the biologically network.

System biology considers that biological functions are not the result of a single gene or protein, but the interaction result of multiple biological molecules with each other. With the development of system biology, the biological network has become a powerful tool to research the complex biological activity. Biological networks can simultaneously study the interaction relationships between different biological molecules. System biology can help us understand the exercise process of biological function and explore the underlying mechanisms of biological processes.

Because changes in biological function are the result of molecular interactions, the functional annotation of differentially expressed genes should consider not only the effects of differentially expressed genes on the pathway, but also the effects of gene interaction on different pathways. In addition, consideration should also be given to the association between pathways.

So, in this paper, we proposed a network-based pathway analysis method. We find the disease related pathways through the analysis of the constructed network. The specific process for the construction of networks includes the following. First, we integrated protein-protein interaction (PPI) information and gene expression profile data into the pathway, and then the candidate genes associated with disease phenotypes were screened using mutual information calculations. Secondly, we integrated gene's GO information into pathway to calculate the correlation between the pathways and then construct the pathway network. Finally, the critical pathways are identified in the network. The experimental results of this method with breast cancer data show that our method can not only find the high risk of gene and signaling pathways, but also find an association between the risk pathways.

2. Methods

Figure 1 describes our network-based pathway analysis method. First, the differentially expressed genes were identified by comparing the disease samples and normal samples, and then they were projected onto protein interaction data. If two protein nodes all appear in differentially expressed genes, they are preserved, and the candidate genes were screened out. Secondly, the biological signaling pathways are obtained from database MSigDB which includes 1329 sets of biological metabolism and signaling pathways. The candidate genes are projected into the pathway, and we calculate the active score of each pathway for every sample according to the document [8]. Then according to the activity vectors of each biological pathway, combining the phenotypic information of the samples, the mutual information between the activity vectors and the phenotypic vectors of the samples is calculated. Next, the semantic similarity of differentially expressed genes in each pathway is calculated using gene's GO knowledge, and we calculate the correlation between the pathways and then construct the pathway network. At last, we used GeneRank algorithm to find the critical pathways in the network. In the following, the detailed explanations of our proposed method are described.

2.1. Project the Candidate Gene to the Pathway. Gene expression data often aim to identify genes that are differentially regulated across different classes of samples, for example, finding the genes affected by a treatment, or finding marker genes that discriminate diseased from healthy subjects. Using gene expression profiles obtained from a number of genes for several samples or experimental conditions, we can obtain a gene set that shows a differential expression pattern across different samples. However, a differential expression gene set does not guarantee the existence of a real interaction between the corresponding proteins. Instead, it only suggests that there may be an interaction between the proteins. To accurately describe the change in gene interactions for several samples or experimental conditions, here we screen genes with PPI network.

In the PPI network, if the two nodes (gene) of the edge are both in the set of differentially expressed genes, then the two genes were reserved; otherwise they were removed. So we get the candidate gene set (CG set). The biological pathway is a complete metabolic pathway which contains all genes that constitute a set. The gene-gene interaction of pathway is different in different tissues or samples. These differences may be caused by genes interactions of the pathway or neighbor pathway. To find the related biological pathway of the disease, we analyze the expression of each gene set in different sample, calculating the activity score of the biological pathway (active fraction) according to [8]; main procedures are as follows and shown in Figure 2.

(1) We get the PPI data from HPRD (human protein reference database) and gene expression data of the breast cancer from NCBI GEO (the Gene Expression Omnibus at the National Center for Biotechnology Information (NCBI)).

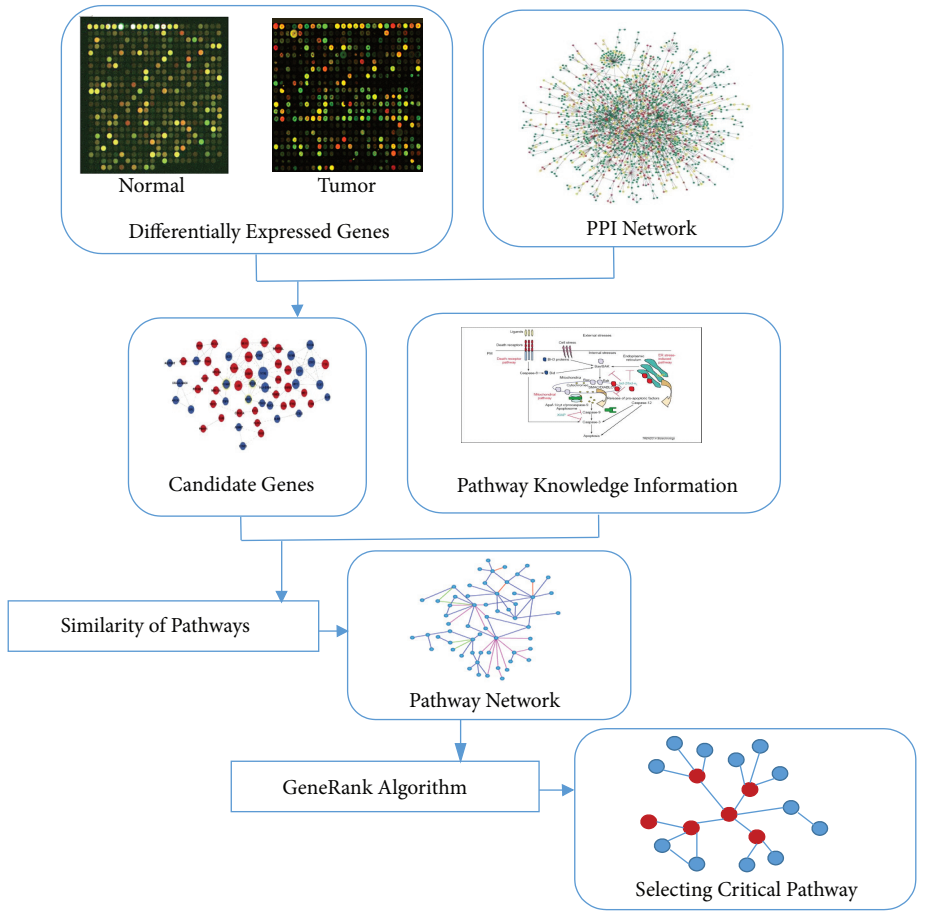


Figure 1: Flow chart of analysis method.

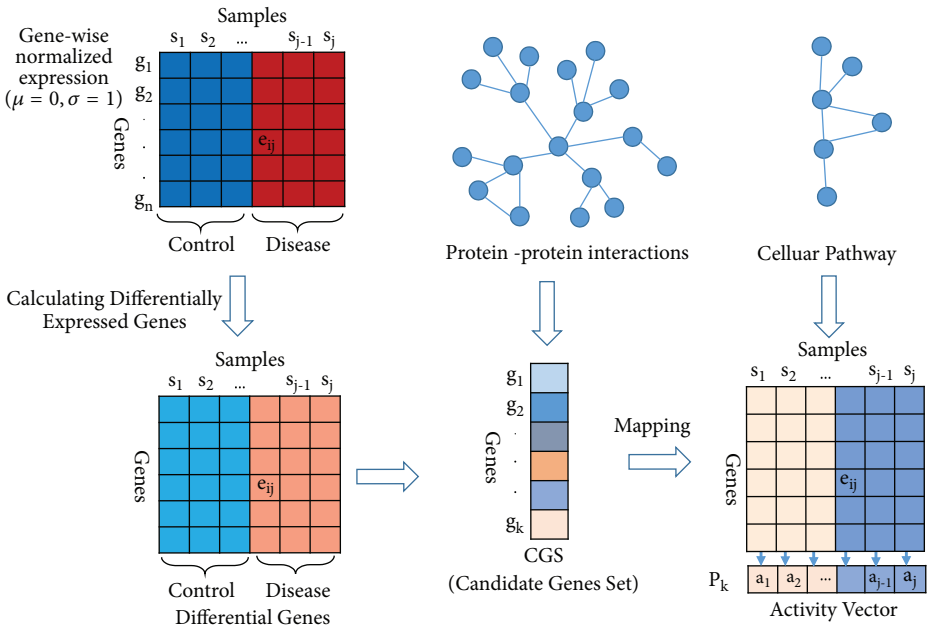


Figure 2: Project the candidate gene to the pathway.

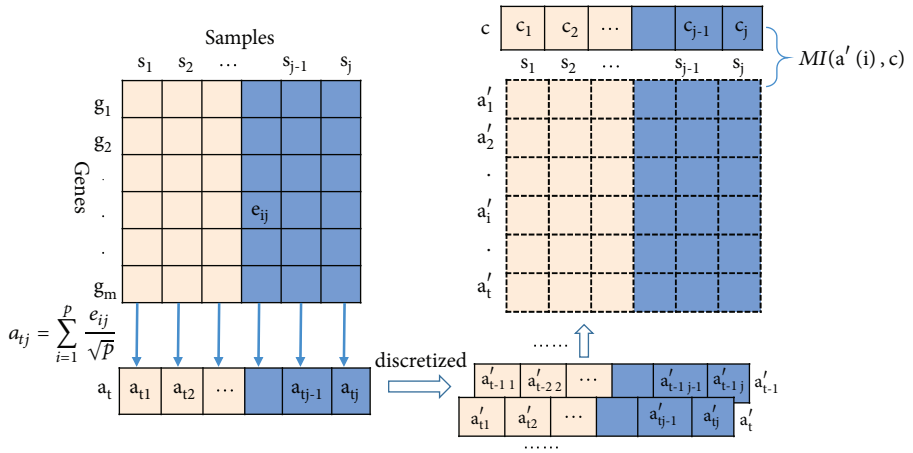


Figure 3: Calculation of mutual information.

(2) We compute differential expression genes in samples and filter through PPI network to get the candidate gene set (CGS).

(3) The biological signaling pathways are obtained from database MSigDB c2 (canonical pathway) which includes 1329 sets of biological metabolism and signaling pathways. The candidate genes are projected into the pathway, calculating the active score of each pathway.

(4) In the GEO data, we divide the samples into disease and control parts and normalize gene expression profiles of the samples and then calculate to get the differentially expressed genes. We map differential gene to the PPI network to obtain the candidate gene set (CGS). The activity scores are calculated according to the expression value of candidate gene set (CGS) on each sample. Gene expression value of gene in biological pathway for every sample constitutes an expression value matrix; each column represents a sample and each row represents a gene. A biological pathway includes p genes whose expression values can form a matrix of p rows. Each biological path corresponds to an activity vector, and the dimension of the vector is the number of samples; that is, in each sample j, an activity score can be calculated. The calculation formula is as follows:

$$a_{tj} = \sum_{i=1}^{p} \frac{e_{ij}}{\sqrt{p}} \tag{1}$$

a_{tj} represents the activity score of the t biological pathway in the j sample, and e_{ij} represents the gene expression value of the i gene in the t biological pathway of the sample j, and p is the number of genes in the biological pathway. After that, we get an activity vector of biological pathways; vector $[a_{t1}, a_{t2}, a_{t3}, \ldots a_{tm}]$ represents the activity score of the t biological pathway in m samples.

Next, a phenotypic vector is constructed based on the phenotypic labels of the samples, and then the mutual information between the activity vectors and the phenotypic vectors of the samples is computed by combining the activity vectors of each biological pathway.

2.2. Mutual Information With Phenotype. Mutual information (MI) is a commonly used method of information measurement in information theory. In [8], the rationale behind using MI to classify cancer patients is explained, and the processing is shown in Figure 3. By calculating the mutual information between the activity vectors and the phenotypic vectors of biological pathways, the correlations between the two vectors are measured, that is, the influence of a biological pathway on the phenotype of the disease.

Constructing a phenotype vector based on the phenotype of the sample, $[c1, c2, c3, \ldots cm]$, the phenotypic vector is a zero-one vector, and if the sample is tumor, the corresponding value is one, otherwise zero.

Using $a(i)$ to indicate the activity score of the i biological pathway on each sample, $a(i)=[a_{i1}, a_{i2}, a_{i3}, \ldots a_{im}]$. c is used to represent the phenotypic vectors of m samples, $c = [c_1, c_2, c_3, \ldots c_m]$. So the correlation between biological pathways and disease phenotype $S(i)$ can be represented by mutual information $MI(a'(i), c)$ between $a'(i)$ vector and c vector. a' is a discretized form of a. The formula is as follows:

$$\begin{aligned} M(i) &= MI\left(a'(i), c\right) \\ &= \sum_{x \in a'} \sum_{y \in c} p(x, y) \log \frac{p(x, y)}{p(x)\, p(y)} \end{aligned} \tag{2}$$

The activity score a is discretized into $\lfloor \log 2(sum\ of\ smaples) + 1 \rfloor$ equally spaced bins to obtain a', respectively, $p(x, y)$ is the joint probability density function of a' and c, and $p(x)$ and $p(y)$ are the marginal probability density function of a' and c.

3. Constructing Pathway Network Related To Disease

The interaction of the pathway can be represented as a network which was constructed as follows.

Let $G(V, E)$ comprise a set V of pathways and a set E denote the weighted pathway-pathway interaction network with $E \in V * V$. Here, we use similarity to define weight of the network. $V = \{p_1, p_2 \ldots p_n\}$, $n = |V|$ is the number of pathways. Matrix A represents the weighted $n \times n$ adjacency

matrix of G, where w_{ij} denotes the weight of the edge connecting pathway p_i to p_j and w_{ij} is calculated by the similarity of p_i and p_j. Supposing two pathways $p_1$and p_2, p_1, p_2 is a set of genes. $p_1 = \{g_{11}, g_{12} \dots g_{1n}\}$, and $p_2 = \{g_{21}, g_{22} \dots g_{2m}\}$.The similarity of p_1 and p_2 can be calculated as weight of network. The similarity between the two pathways is the ratio of the sum of the similarities of all similar genes to the sum of the two pathway's elements. The formula is shown in

$$sim_{pi,pj} = \frac{\sum_{g_{ix} \in pi, g_{jy} \in pj} sim\left(g_{ix}, g_{jy}\right)}{M + N} \tag{3}$$

$M = |p_i|$ is the number of genes on p_i, and $N = |p_j|$ is the number of genes on p_j.

By this formula, the similarity values of each pair of candidate pathways can be calculated, and the pathway network can be constructed to identify the critical pathway.

We use GOSemSim package of R language [9] to calculate the semantic similarity ($sim(g_{ix}, g_{jy})$) between genes, including molecular function similarity (MF), biological processes similarity (BP), and Cellular Component (CC) similarity. Since a pathway contains multiple genes, we calculate the similarity of genes on the pathway as the similarity between the pathways to obtain the pathway similarity matrix $sim_{pi,pj}$.

4. Identifying Significant Pathways of Pathway Network

Based on the biological pathway information and the similarity information of genes in the pathway, we constructed the biological pathway interaction network. In the network, each node represents a biological pathway, and if the two pathways contain differentially expressed genes, then they are connected. And the similarity of the differential genes on the two pathways is used as the similarity of the two pathways.

After constructing the biological pathway network, we hope to find biological pathways related to cancer phenotype. We use the random walk algorithm combined with phenotypic information to find the key nodes in the pathway network. The GeneRank method is a sorting algorithm proposed by Morrison et al. [10]. In our method, the initial node order is determined by the value of mutual information MI, and the transfer matrix is obtained by the pathway similarity matrix ($sim_{pi,pj}$). According to the random walk GeneRank algorithm, starting from the initial node, the final stable pathway node sequence is obtained by iterative calculation of the transfer matrix. This approach considers both the pathway and phenotype information and the semantic similarity between pathways, thus avoiding the node as isolated individuals and ignoring the important nodes that are highly correlated with other nodes. Therefore, this method can find potential pathway nodes.

We let set $P = \{p_1, p_2, \dots, p_N\}$ represent n nodes in the pathway network. According to whether there is a similarity relationship between the two pathway nodes, the adjacency matrix $w(i, j)$ can be obtained by a similar matrix $sim_{pi,pj}$ according to a threshold θ, as shown in

$$w(i, j) = \begin{cases} 1 & \text{if } \mathrm{sim}(\mathrm{pi}, \mathrm{pj}) \geq \theta \\ 0 & others \end{cases} \tag{4}$$

Sort the pathway set P according to the mutual information (MI) of pathway and phenotype. The initial row rank of P is obtained, which is denoted as x, $ex = (ex_1, ex_2, \dots, ex_N)$.

According to the definition, W is a symmetric matrix, $w_{i,j} = W(i, j) = W(j, i) = w_{j,i}$. According to graph theory, the degree of the i node is equal to the sum of the elements of row i of matrix w and is expressed in $\deg_i$, as shown in

$$deg_i = \sum_{j=1}^{n} w_{i,j} = \sum_{j=1}^{n} w_{j,i} \tag{5}$$

d is the damping coefficient, and the closer the value of d is to 0, the greater the impact of the mutual information is on the node sorting; on the contrary, if d is closer to 1, node sorting is more affected by the similarity. In our algorithm, we initialized d to 0.85.

The algorithm steps are as follows:

Input: initialization vector $ex = (ex_1, ex_2, \dots, ex_N)$; adjacency matrix w; parameter d.

ε is the error value; max is the maximum iterations.

Output: r;

(1) Data Preprocessing:

$$r^{[0]} = \frac{ex}{\|ex\|_1};$$
$$r^{[n]} = \left(\mathrm{r}_1^{[n]}, \mathrm{r}_2^{[n]}, \dots, \mathrm{r}_N^{[n]}\right); \quad n = 0, 1, 2, \dots. \tag{6}$$

(2) Iteration:

$$\mathrm{r}_j^{[n]} = (1 - d)\, ex_j + d \sum_{i=1}^{N} \frac{w_{ij} r_i^{[n-1]}}{deg_i}, \quad 1 \leq j \leq N;$$
$$res^{[n]} = \left\| r^{[n]} - r^{[n-1]} \right\|_1; \tag{7}$$

(3) Stop Condition

If $res^{[n]} \leq \varepsilon$ *or* $n \geq max$, *Stop*;

Else goto (2);

Return r=$r^{[n]}$;

5. Experiments and Results

5.1. Data. The Breast Cancer Dataset was downloaded from GEO (Gene Expression Omnibus) website (https://www.ncbi.nlm.nih.gov/geo/), including GSE33447 [11], GSE9309, GSE15852[12], GSE5364 [13]. and GSE20437 [14]. The dataset consists of 484 samples obtained from comparing 387 breast cancer samples with 97 normal samples, as shown in Table 1.

The gene expression profiles of 369 cases of breast cancer and 73 cases of normal breast tissues were obtained, and the differentially expressed genes were analyzed, PPI network was obtained from the Human Protein Reference

Table 1: Breast Cancer Data Set.

DataSet	Normal	Tumor
GSE 9309	9	132
GSE 15852	43	43
GSE 5364	13	186
GSE 33447	8	8
GSE 20437	24	18

Table 2: Top 15 pathways identified with our method.

Rank	Pathway Name	gene	p-value
1	KEGG_FATTY_ACID_METABOLISM	42	5.76e-12
2	KEGG_STARCH_AND_SUCROSE_METABOLISM	52	0.0047780
3	KEGG_SPLICEOSOME	128	3.63e-12
4	KEGG_PPAR_SIGNALING_PATHWAY	69	8.64e-21
5	KEGG_P53_SIGNALING_PATHWAY	69	0.0005045
6	KEGG_ADIPOCYTOKINE_SIGNALING_PATHWAY	68	2.23e-09
7	PID_SHP2_PATHWAY	58	2.40e-05
8	PID_BARD1_PATHWAY	29	0.0010373
9	REACTOME_MRNA_3_END_PROCESSING	36	0.0137007
10	REACTOME_METABOLISM_OF_NON_CODING_RNA	49	0.0008123
11	REACTOME_CELL_CYCLE	421	1.02e-08
12	REACTOME_SIGNALING_BY_BMP	23	0.0247461
13	REACTOME_TRANSPORT_OF_MATURE_TRANSCRIPT_TO_CYTOPLASM	54	0.0426007
14	REACTOME_CELL_CYCLE_MITOTIC	325	9.35e-07
15	REACTOME_PROCESSING_OF_CAPPED_INTRONLESS_PRE_MRNA	23	0.0127461

Database (http://www.hprd.org/) [15]. The pathways were downloaded from the Molecular Signatures Database website (http://software.broadinstitute.org/gsea/msigdb) [16]. The database mainly collects gene set which was annotated with certain biological functions. We chose C2 gene sets (canonical pathways, 1392 gene sets); the set of genes is derived from several major biological pathway databases, including BioCarta [17], KEGG [18], and Reactome [19] databases.

To identify the significance of the given pathway, first, we computed differential expression genes of the sample datasets (GSE9309, GSE15852) and then dealt with the PPI data. PPI network was mapped into the differential expression gene; we obtained candidate gene set (CGS). Secondly, the candidate genes are mapped into the pathways, calculating the active score of each pathway, and then the mutual information is computed by combining the activity vectors of each biological pathway. Finally, we calculate the similarity of the pathway to obtain pathway network, and the random walk algorithm combined with phenotypic information was used to find significance node of the pathway network.

To provide a more comprehensive understanding of the proposed method, we discuss the method from the following aspects separately.

5.2. The Results of Pathway Recognition. According to the above description, the rank of each pathway is the degree of relevance between the given pathway and the corresponding disease. The rank is calculated by algorithm (see in Section 4). In this study, the significance of the pathway was tested by hypergeometric distribution of each pathway with annotated differential genes using formula (8). The top 15 pathways with p-value are shown in Table 2; we selected part of the interaction pathways, as shown in Figure 4.

$$p - value = 1 - \sum_{i=0}^{x-1} \frac{\binom{M}{i}\binom{U-M}{N-i}}{\binom{U}{N}} \tag{8}$$

U—the number of genes in the human genome;

N—the number of differential genes;

M—the number of genes in the pathway;

x—the number of differentially expressed genes in this pathway.

The top 1 pathway is KEGG_FATTY_ACID_METABOLISM; the pathway is supported by [20–22], and the KEGG_SPLICEOSOME [23–25], REACTOME_ METABOLISM_OF_NON_CODING_RNA [26, 27], and REACTOME_CELL_CYCLE [28] are all very important metabolic pathways, and we focus on the following important pathways.

One significant pathway identified by our method was P53_SIGNALING_PATH-WAY [29–31]. Currently, breast cancer is the most prevalent cancer diagnosed in women, with an estimated 1.8 million cases reported worldwide in 2013 [32, 33]. Radiation is commonly adopted as an adjuvant therapy for the management of breast cancer [34]. However,

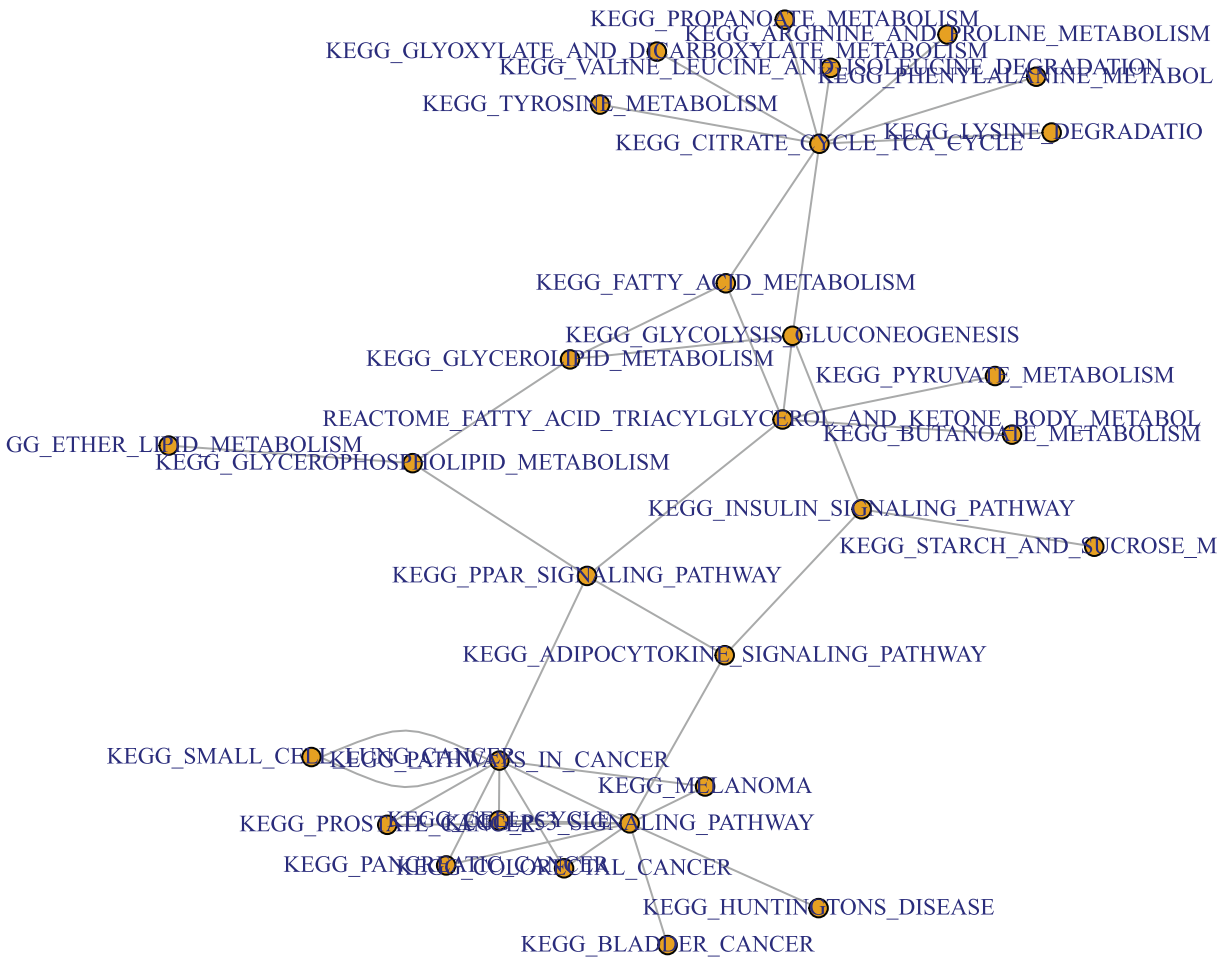


Figure 4: Interaction between disease pathways.

there is growing evidence that autophagy is induced by ionizing radiation, and this induction plays a crucial role in radiosensitivity [35, 36]. Furthermore, the regulatory effect of autophagy in radiation-induced cell death remains controversial, and the underlying molecular mechanisms remain to be fully characterized. The research of [37] shows that the p53/DRAM signaling pathway appears to contribute to radiation-induced autophagic cell death in MCF-7 breast cancer cells.

Another significant pathway was KEGG_PPAR_SIGNALING_PATHWAY. In [38], the authors analyzed six pathological complete response (pCR) patients and 25 patients with non-pCR; 300 probes (231 genes) were identified as differentially expressed between pCR and residual disease by the SAM program when the fold change was >2. The gene functional enrichment analysis revealed 15 prominent gene categories that were different between pCR and non-pCR patients, most notably the genes involved in the peroxisome proliferator-activated receptor (PPAR), DNA repair, and ER signal pathways and in the immune-related gene cluster; they believe that the PPAR pathway may be an important predictor of genes that are involved in the chemotherapy response.

The other pathway was PID_SHP2_PATHWAY; in [39], the researchers show a fundamental role for Src-homology 2 domain-containing phosphatase 2 (SHP2) in these processes in human epidermal growth factor receptor 2- (HER2-) positive and triple-negative breast cancers. Knockdown of SHP2 eradicated breast tumor-initiating cells in xenograft models, and SHP2 depletion also prevented invasion in three-dimensional cultures and in a transductal invasion assay in vivo. Notably, SHP2 knockdown in established breast tumors blocked their growth and reduced metastasis. Mechanistically, SHP2 activated stemness-associated transcription factors, including v-myc myelocytomatosis viral oncogene homolog (c-Myc) and zinc finger E-box binding homeobox 1 (ZEB1), which resulted in the repression of let-7 microRNA and the expression of a set of "SHP2 signature" genes. They found these genes to be simultaneously activated in a large subset of human primary breast tumors that are associated with invasive behavior and poor prognosis. These results provide new insights into the signaling cascades influencing tumor-initiating cells as well as a rationale for targeting SHP2 in breast cancer.

Moreover, we obtained top 15 pathways; their p-values are all less than 0.05. In order to test the effectiveness of this

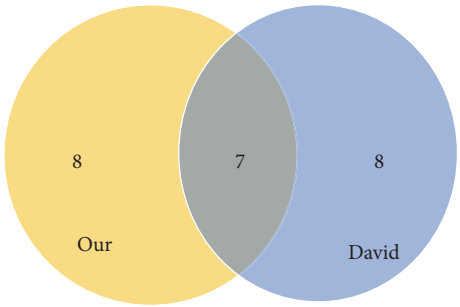


Figure 5: DAVID method and our method to identify the risk pathway.

Table 3: Common pathways with our method and David.

Pathway Name	p-value
KEGG_FATTY_ACID_METABOLISM	5.76e-12
KEGG_SPLICEOSOME	3.63e-12
KEGG_PPAR_SIGNALING_PATHWAY	8.64e-21
KEGG_P53_SIGNALING_PATHWAY	0.0005045
KEGG_ADIPOCYTOKINE_SIGNALING_PATHWAY	2.23e-09
REACTOME_CELL_CYCLE	1.02e-08
REACTOME_CELL_CYCLE_MITOTIC	9.35e-07

method, we compare with the most used DAVID software. The identified top 15 pathways are compared with DAVID method and our method to identify the risk pathway of breast cancer as shown in Figure 5; the common pathways are shown in Table 3.

DAVID is a widely used and approved method. According to the pathway set we have identified (in Tables 2 and 5). Pathways identified by our method and DAVID method have obvious intersection (in Figure 5); the p-value of the pathway is also significant. In addition, compared with DAVID, we also found some pathways which DAVID did not find. In Table 2, PID_BARD1_PATHWAY [40–45], PID_SHP2_PATHWAY [39], REACTOME_TRANSPORT_OF_MATURE_TRANSCRIPT_TO_CYTOPLASM [46], and REACTOME_METABOLISM_OF_NON_CODING_RNA [46, 47] are the pathways that our method identified but DAVID did not. The relationship between them and breast cancer has been evaluated in detail and is affirmative in the corresponding literature.

So, according to the above analysis, it can be concluded that our proposed method is effective in identifying the important pathways of the complex diseases.

5.3. Test on Dataset

5.3.1. Results and Analysis. To estimate the classification performance, firstly we prepared our dataset (GSE9309, GSE15852) and took 80 genes in the selected pathways as features, and SVM [48] is employed to classify the selected samples. Next, a 10-fold cross-validation was used to train and test SVM. The above experiment was repeated 100 times; the average value of the 100-time calculation is taken as the final result. In order to evaluate our method, we compared the classification results of cancer and normal samples with the commonly used methods based on differential expression. We choose the T-test method, and the T-test can be used to test whether the means of two independent normal distribution samples are equal. For gene expression data, we can test whether there is a significant difference in the expression of a gene between different phenotypes, that is, to identify differentially expressed genes through tests. The results obtained by the T-test method are compared with our method according to the genes contained in the biological pathway, and the same number of genes is sorted according to the order. The comparison results show that the method is superior to the T-test method, and the experimental results show that our proposed method is more effective in distinguishing cancer from normal samples, and the results are shown in Figure 6.

We applied the feature gene set to the independent gene expression datasets (GSE5364, GSE33447, and GSE20437). This set of data is not related to the data previously used, and they are independent of each other. In order to evaluate our proposed method objectively, we used the same number of biological pathways or gene markers to compare the results. The biological pathway markers obtained by our method have a good discrimination in the dataset, higher than 0.6, close to 0.7, and the results are shown in Figure 7. However, the AUC of the currently reported prognostic models on independent datasets is very difficult to reach 0.7 [49], which shows the superiority of our method.

5.3.2. Test on Other Datasets. Secondly, in order to test the robustness of our method, we apply the method in this article to four gastric cancer gene expression datasets. The dataset was also downloaded from GEO, which includes GSE63089 [50], GSE56807 [51], GSE33335 [52], and GSE19826 [53]. PPI data was obtained from STRING (https://stringdb-static.org/); the sources of other data are the same. The dataset consists of 177 samples obtained from comparing 87 tumor samples with 90 normal samples, which is shown in Table 4; the top 15 pathways with p-value are shown in Table 5.

We used our method to test on the gastric cancer dataset, and we selected the same number of biological pathways and gene markers. Compared with the T-test, our method has also achieved a higher or comparable value. We used GSE63089 and GSE56807 as the training set, with GSE33335 and GSE19826 as the test set. A 10-fold cross-validation was used to train and test SVM. The above experiment was repeated 50 times; the average value is taken as the final result. Our method also has good discrimination in this dataset, higher than 0.6 and better than T-test; the results on gastric cancer data are shown in Figure 8.

5.4. Conclusions. Complex diseases, especially cancer, are extremely harmful to human health. Therefore, the identification of cancer markers is the key of the study. Pathway analysis combined with multiple types of high-throughput data reflects the biological processes more clearly. Therefore, pathway-based complex disease analysis method has become

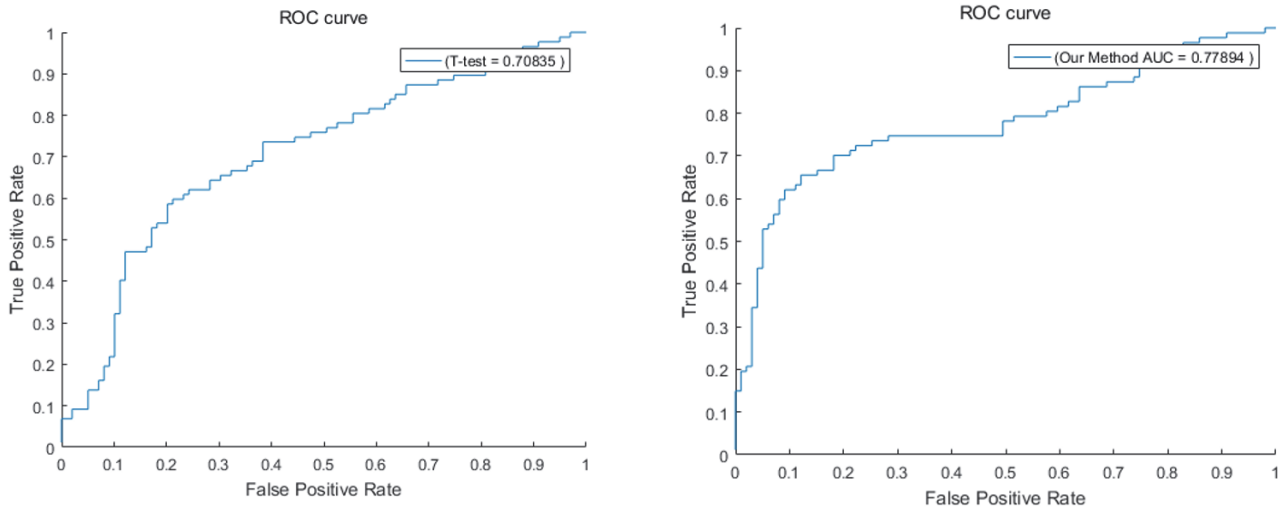


Figure 6: Test on dataset.

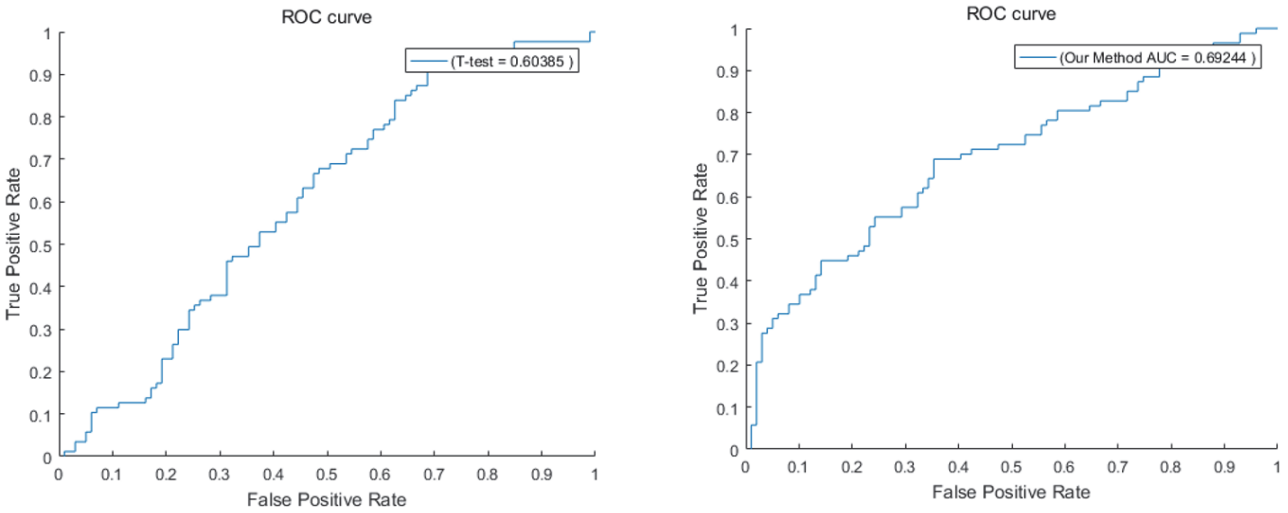


Figure 7: Test on independent datasets.

a hot research topic. Unlike previously available pathway analysis methods, we have considered not only the genes interaction of the pathway, but also the interaction between the pathways.

In this paper, we proposed a new approach to consider the correlation between biological pathways and establish a biological pathway interaction network. Then, the GeneRank algorithm based on random walk and the mutual information of phenotype were used to select cancer related biological pathways. Finally, we use the support vector machine and feature selection method to apply to cancer datasets. The results show that our method achieves better results than T-test method. In addition, the validation in the independent dataset and the functional analysis of the biological pathway indicate that the pathway we identified as a biological marker of disease is more accurate and reliable. We will employ more datasets to assess the validity of our approach in future research.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This study is supported by the Natural Science Foundation of Hubei Province of China (Program no. 2015CFB524 and Program no. 2016CKB705), the Fundamental Research Funds for the Central Universities (Program no. 2015BQ023 and Program no. 2014QC008), and the Students Research Fund

Table 4: Gastric Cancer Data Set.

DataSet	Normal	Tumor
GSE 63089	45	45
GSE 56807	5	5
GSE 33335	25	25
GSE 19826	15	12

Table 5: Top 15 pathways identified with our method in Gastric Cancer Data Set.

Rank	Pathway Name	gene	p-value
1	KEGG_DNA_REPLICATION	36	1.50e-20
2	KEGG_PROTEASOME	48	9.07e-05
3	KEGG_ARGININE_AND_PROLINE_METABOLISM	54	0.000489
4	KEGG_GLYCEROLIPID_METABOLISM	49	5.12e-05
5	KEGG_PURINE_METABOLISM	159	1.68e-12
6	KEGG_PATHWAYS_IN_CANCER	328	2.52e-09
7	KEGG_SPLICEOSOME	128	0.0003348
8	KEGG_NUCLEOTIDE_EXCISION_REPAIR	44	0.0001406
9	KEGG_MELANOGENESIS	102	0.0481885
10	KEGG_RNA_DEGRADATION	59	0.0003626
11	KEGG_MAPK_SIGNALING_PATHWAY	267	0.0014066
12	KEGG_GLYCEROLIPID_METABOLISM	49	5.119e-05
13	KEGG_DRUG_METABOLISM_CYTOCHROME_P450	72	0.0115133
14	KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION	267	0.0183417
15	KEGG_PROSTATE_CANCER	89	0.0274642

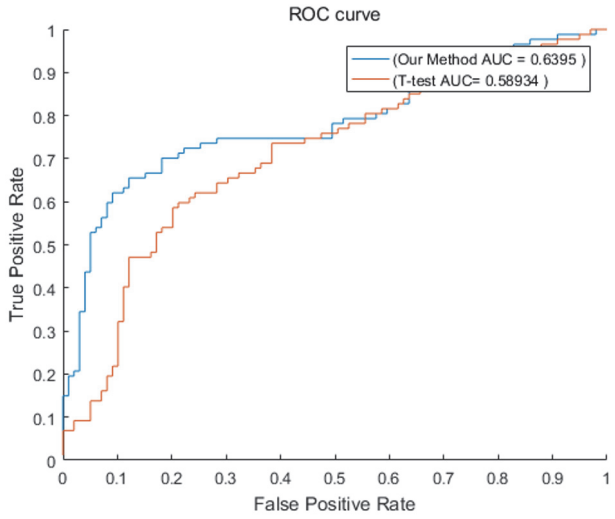


Figure 8: Test on independent gastric cancer dataset.

(SRF) of Huazhong Agricultural University (Program no. 2017400).

References

[1] L. Cowen, T. Ideker, B. J. Raphael, and R. Sharan, "Network propagation: A universal amplifier of genetic associations," *Nature Reviews Genetics*, vol. 18, no. 9, pp. 551–562, 2017.

[2] U. D. Akavia, O. Litvin, J. Kim et al., "An integrated approach to uncover drivers of cancer," *Cell*, vol. 143, no. 6, pp. 1005–1017, 2010.

[3] Q. Zhang, J. Li, H. Xue, L. Kong, and Y. Wang, "Network-based methods for identifying critical pathways of complex diseases: A survey," *Molecular BioSystems*, vol. 12, no. 4, pp. 1082–1089, 2016.

[4] C. S. Greene and B. F. Voight, "Pathway and network-based strategies to translate genetic discoveries into effective therapies," *Human Molecular Genetics*, vol. 25, no. 2, pp. R94–R98, 2016.

[5] A. Subramanian, P. Tamayo, V. K. Mootha et al., "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15545–15550, 2005.

[6] A. L. Tarca, S. Draghici, P. Khatri et al., "A novel signaling pathway impact analysis," *Bioinformatics*, vol. 25, no. 1, pp. 75–82, 2009.

[7] https://david.ncifcrf.gov/.

[8] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis," *Molecular Systems Biology*, vol. 3, no. 1, article 140, 2007.

[9] G. Yu, F. Li, Y. Qin, X. Bo, Y. Wu, and S. Wang, "GOSemSim: An R package for measuring semantic similarity among GO terms and gene products," *Bioinformatics*, vol. 26, no. 7, Article ID btq064, pp. 976–978, 2010.

[10] J. L. Morrison and etal., "GeneRank: Using search engine technology for the analysis of microarray experiments," *Bmc Bioinformatics*, vol. 6, pp. 1–14, 2005.

[11] Z.-Q. Lian, Q. Wang, W.-P. Li, A.-Q. Zhang, and L. Wu, "Screening of significantly hypermethylated genes in breast cancer using microarray-based methylated-CpG island recovery assay and identification of their expression levels," *International Journal of Oncology*, vol. 41, no. 2, pp. 629–638, 2012.

[12] I. B. Pau Ni, "Gene expression patterns distinguish breast carcinomas from normal breast tissues: the Malaysian context," *Pathology Research & Practice*, vol. 206, pp. 223–228, 2010.

[13] K. Yu, K. Ganesan, L. K. Tan et al., "A precisely regulated gene expression cassette potently modulates metastasis and survival in multiple solid cancers," *PLoS Genetics*, vol. 4, no. 7, Article ID e1000129, 2008.

[14] K. Graham, A. De Las Morenas, A. Tripathi et al., "Gene expression in histologically normal epithelium from breast cancer patients and from cancer-free prophylactic mastectomy patients shares a similar profile," *British Journal of Cancer*, vol. 102, no. 8, pp. 1284–1293, 2010.

[15] B. Liu and B. Hu, "HPRD: a high performance RDF database," *International Journal of Parallel, Emergent and Distributed Systems*, vol. 25, no. 2, pp. 123–133, 2010.

[16] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov, "Molecular signatures database (MSigDB) 3.0," *Bioinformatics*, vol. 27, no. 12, pp. 1739-1740, 2011.

[17] http://www.biocarta.com/.

[18] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa, "KEGG for representation and analysis of molecular networks involving diseases and drugs," *Nucleic Acids Research*, vol. 38, no. 1, pp. D355–D360, 2009.

[19] E. Schmidt, E. Birney, D. Croft et al., "Reactome – A Knowledgebase of Biological Pathways," in *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops*, vol. 4277 of *Lecture Notes in Computer Science*, pp. 710–719, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[20] A. Huqi, "Cancer and inhibition of fatty acid oxidation," *Heart and Metabolism*, no. 51, pp. 27–30, 2011.

[21] W. Zhou, Y. Tu, P. J. Simpson, and F. P. Kuhajda, "Malonyl-CoA decarboxylase inhibition is selectively cytotoxic to human breast cancer cells," *Oncogene*, vol. 28, no. 33, pp. 2979–2987, 2009.

[22] J. N. Thupari, M. L. Pinn, and F. P. Kuhajda, "Fatty acid synthase inhibition in human breast cancer cells leads to malonyl-CoA-induced inhibition of fatty acid oxidation and cytotoxicity," *Biochemical and Biophysical Research Communications*, vol. 285, no. 2, pp. 217–223, 2001.

[23] M. C. Wahl, C. L. Will, and R. Lührmann, "The spliceosome: design principles of a dynamic RNP machine," *Cell*, vol. 136, no. 4, pp. 701–718, 2009.

[24] H. Le Hir, E. Izaurralde, L. E. Maquat, and M. J. Moore, "The spliceosome deposits multiple proteins 20-24 nucleotides upstream of mRNA exon-exon junctions," *EMBO Journal*, vol. 19, no. 24, pp. 6860–6869, 2000.

[25] J. P. Staley and C. Guthrie, "Mechanical devices of the spliceosome: Motors, clocks, springs, and things," *Cell*, vol. 92, no. 3, pp. 315–326, 1998.

[26] T. R. Mercer, M. E. Dinger, and J. S. Mattick, "Long non-coding RNAs: insights into functions," *Nature Reviews Genetics*, vol. 10, no. 3, pp. 155–159, 2009.

[27] S. Sarkar, H. Dubaybo, S. Ali, P. Goncalves, and S. L. Kollepara, "Long non-coding RNAs (lncRNAs) and viral infections," *Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie*, vol. 3, pp. 465–477, 2013.

[28] T. Mazza and M. Cavaliere, "Cell Cycle and Tumor Growth in Membrane Systems with Peripheral Proteins," *Electronic Notes in Theoretical Computer Science*, vol. 227, no. C, pp. 127–141, 2009.

[29] Y. N. Lee, Y. G. Park, Y. H. Choi, and Y. S. Cho, "CRE-transcription factor decoy oligonucleotide inhibition of MCF-7 breast cancer cells: Cross-talk with p53 signaling pathway," *Biochemistry*, vol. 39, no. 16, pp. 4863–4868, 2000.

[30] O. E. C. Mg, "Differential regulation of signal transduction pathways in wild type and mutated p53 breast cancer epithelial cells by copper and zinc," *Archives of Biochemistry & Biophysics*, vol. 423, no. 2, pp. 351–361, 2004.

[31] J. H. Seo, "PEGylated conjugated linoleic acid stimulation of apoptosis via a p53-mediated signaling pathway in MCF-7 breast cancer cells," *European Journal of Pharmaceutics & Biopharmaceutics Official Journal of Arbeitsgemeinschaft Für Pharmazeutische Verfahrenstechnik E V*, vol. 70, no. 621, 2008.

[32] K. N. Dalby, I. Tekedereli, G. Lopez-Berestein, and B. Ozpolat, "Targeting the prodeath and prosurvival functions of autophagy as novel therapeutic strategies in cancer," *Autophagy*, vol. 6, no. 3, pp. 322–329, 2010.

[33] C. Fitzmaurice, D. Dicker, A. Pain et al., "The global burden of cancer 2013," *JAMA Oncology*, vol. 1, pp. 505–527, 2015.

[34] P. G. Tsoutsou, M. I. Koukourakis, D. Azria, and Y. Belkacémi, "Optimal timing for adjuvant radiation therapy in breast cancer. A comprehensive review and perspectives," *Critical Review in Oncology/Hematology*, vol. 71, no. 2, pp. 102–116, 2009.

[35] Y. Yang, "Autophagy and its function in radiosensitivity," *Tumour Biology the Journal of the International Society for Oncodevelopmental Biology & Medicine*, vol. 36, pp. 1–9, 2015.

[36] D. A. Gewirtz, "The four faces of autophagy: implications for cancer therapy," *Cancer Research*, vol. 74, no. 3, pp. 647–651, 2014.

[37] L. Cui, Z. Song, B. Liang, L. Jia, S. Ma, and X. Liu, "Radiation induces autophagic cell death via the p53/DRAM signaling pathway in breast cancer cells," *Oncology Reports*, vol. 35, no. 6, pp. 3639–3647, 2016.

[38] Y. Z. Chen, "PPAR signaling pathway may be an important predictor of breast cancer response to neoadjuvant chemotherapy," *Cancer Chemotherapy & Pharmacology*, vol. 70, 2012.

[39] X. Wang, "Abstract 2150: The role of SHP2 in HER2+ breast cancer," *Cancer Research*, vol. 72, no. 8, pp. 2150-2150, 2014.

[40] M. Fabbro, K. Savage, K. Hobson et al., "BRCA1-BARD1 complexes are required for P53Ser-15 phosphorylation and a G1/S arrest following ionizing radiation-induced DNA damage," *The Journal of Biological Chemistry*, vol. 279, no. 30, pp. 31251–31258, 2004.

[41] G. Birrane, A. K. Varma, A. Soni, and J. A. A. Ladias, "Crystal structure of the BARD1 BRCT domains," *Biochemistry*, vol. 46, no. 26, pp. 7706–7712, 2007.

[42] R. Baer and T. Ludwig, "The BRCA1/BARD1 heterodimer, a tumor suppressor complex with ubiquitin E3 ligase activity," *Current Opinion in Genetics & Development*, vol. 12, no. 1, pp. 86–91, 2002.

[43] "Bard, BRCA1 associated RING domain 1".

[44] D. Fox III, I. Le Trong, P. Rajagopal, P. S. Brzovic, R. E. Stenkamp, and R. E. Klevit, "Crystal structure of the BARD1 ankyrin repeat domain and its functional consequences," *The Journal of Biological Chemistry*, vol. 283, no. 30, pp. 21179–21186, 2008.

[45] I. Irminger-Finger, W.-C. Leung, J. Li et al., "Identification of BARD1 as mediator between proapoptotic stress and p53-dependent apoptosis," *Molecular Cell*, vol. 8, no. 6, pp. 1255–1266, 2001.

[46] P. Boyle and P. Boffetta, "Alcohol consumption and breast cancer risk," *Breast Cancer Research*, vol. 11, no. S3, 2009.

[47] P. Paci, T. Colombo, and L. Farina, "Computational analysis identifies a sponge interaction network between long non-coding RNAs and messenger RNAs in human breast cancer," *BMC Systems Biology*, vol. 8, article 83, 2014.

[48] C. Chang and C. Lin, "LIBSVM: a Library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, article 27, 2011.

[49] G. Abraham, A. Kowalczyk, S. Loi, I. Haviv, and J. Zobel, "Prediction of breast cancer prognosis using gene set statistics provides signature stability and biological context," *BMC Bioinformatics*, vol. 11, article no. 277, 2010.

[50] X. T. Zhang, Z. H. Ni, Z. P. Duan et al., "Overexpression of E2F mRNAs associated with gastric cancer progression identified by the transcription factor and miRNA co-regulatory network analysis," *PLoS ONE*, vol. 10, no. 2, article e0116979, 2015.

[51] J. Wang, Z. Ni, Z. Duan, G. Wang, and F. Li, "Altered expression of hypoxia-inducible factor-1α (HIF-1α) and its regulatory genes in gastric cancer tissues," *PLoS ONE*, vol. 9, no. 6, Article ID e99835, 2014.

[52] L. Cheng, P. Wang, S. Yang et al., "Identification of genes with a correlation between copy number and expression in gastric cancer," *BMC Medical Genomics*, vol. 5, article 14, 2012.

[53] Q. Wang, Y. G. Wen, D. P. Li et al., "Upregulated INHBA expression is associated with poor survival in gastric cancer," *Medical Oncology*, vol. 29, no. 1, pp. 77–83, 2012.

Hindawi
BioMed Research International
Volume 2018, Article ID 1501430, 10 pages
https://doi.org/10.1155/2018/1501430

Research Article

Genetic Polymorphism Study on *Aedes albopictus* of Different Geographical Regions Based on DNA Barcoding

Yiliang Fang,[1] **Jianqing Zhang**,[1] **Rongquan Wu**,[2] **Baohai Xue**,[1] **Qianqian Qian**,[3] **and Bo Gao**[1]

[1]*Fujian International Travel Healthcare Center, Fuzhou, Fujian 350001, China*
[2]*Quanzhou Entry-Exit Inspection and Quarantine Bureau Comprehensive Technical Service Center, Quanzhou 362000, Fujian Province, China*
[3]*Fujian Medical University, Fuzhou, Fujian 350001, China*

Correspondence should be addressed to Bo Gao; gaobo28083@sina.com

Received 30 December 2017; Accepted 19 March 2018; Published 29 May 2018

Academic Editor: Yudong Cai

Aedes albopictus is a very important vector for pathogens of many infectious diseases including dengue fever. In this study, we explored the genetic polymorphism of *Aedes albopictus* strains in different geographical regions using DNA barcoding of mitochondrial COI (*MT-COI*) gene. We collected *MT-COI* sequence of 106 *Aedes albopictus* mosquitos from 6 provinces in China including Fujian, Guangdong, Hainan, Yunnan, and Taiwan. The length of the sequences is 709bp with the content of A+T (67.7%) greater than that of G+C (32.3%). We identified mutations in 90 (13.68%) loci, of which 57 (63.33%) are transitions, 28 (31.11%) are transversions, and 5 (5.56%) are hypervariable loci. In addition, we obtained 42 haplotypes, 4 (9.52%) of which are shared among different populations. The haplotype diversity of *Aedes albopictus* is 0.882 and nucleotide diversity is 0.01017. Moreover, the pedigree network diagram shows that most haplotypes are under parallel evolution, suggesting a local expansion of *Aedes albopictus* in history. Finally, the Neighbor-Joining tree of *MT-COI* haplotypes reveals a certain correlation between haplotype clusters and geographical distribution, and there are differences among *Aedes albopictus* in different geographical regions. In conclusion, DNA barcoding of *MT-COI* gene is an effective method to study the genetic structure of *Aedes albopictus*.

1. Introduction

Aedes albopictus, belonging to the *Diptera Nematocera Culicidac Aedes* genus, are widely distributed all over the world [1–4]. It was originated in Southeast Asia and spread rapidly into many countries in Africa, the Middle East, Europe, and America in the past three decades [1, 5]. International trade, especially that of used tires at ports, has accelerated its transmission [6]. In China, though *Aedes albopictus* was observed from Hainan island in the south all over to Liaoning province in the north, it was most gathered mostly in the south of 30° north latitude such as the Fujian province [7]. Understanding the population genetic polymorphism of *Aedes albopictus* is critical to its prevention and control.

DNA barcoding is a biometrics based on the mitochondrial cytochrome C oxidase subunit I (*MT-COI*) gene (about 500-600 bp), which was first introduced by Canadian biologist Paul Hebert in 2003 [8, 9]. This technology has received widespread attention since then. Mitochondrial COI gene sequences with strict maternal inheritance, conservative genetic makeup, and moderate evolution rate but higher nuclear DNA characteristics [10] have been widely used in the identification of mammals, fish, insects, birds, and other species [11–15]. However, to our best knowledge, there is no research on genetic polymorphism of *Aedes albopictus* at species level based on DNA barcoding at present. The reason might be due to the high cost in collecting and sequencing *Aedes albopictus* strains in a wide region, which might even be infeasible in a small country.

In this study, we collected and sequenced the *MT-COI* gene of 106 *Aedes albopictus* strains from 6 provinces in China including Fujian, Guangdong, Hainan, Yunnan, and Taiwan. We then performed down-stream bioinformatics and phylogenetic analysis on the sequences. This study reveals

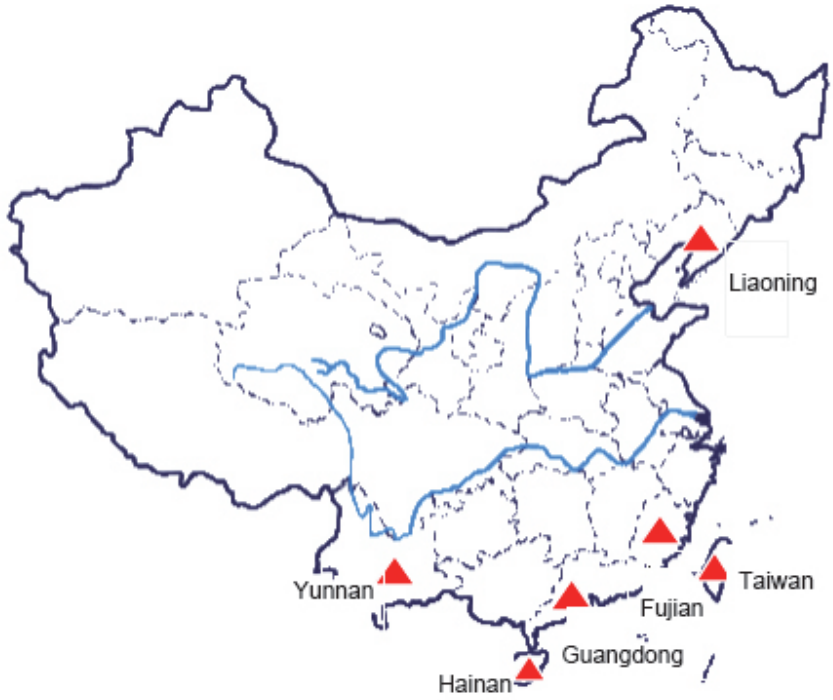


Figure 1: Collecting places of different geographical strains of *Aedes albopictus*.

the genetic polymorphism of *Aedes albopictus* in South China, which is important for the prevention and control of infectious diseases spread by *Aedes albopictus* like dengue fever and yellow fever [16].

2. Materials and Methods

2.1. Experimental Mosquitos. We collected mosquitos from Fujian province (Fuzhou, Zhangzhou, Jiangling, and Wuyishan), Guangdong province (Guangzhou), Hainan province (Diaoluoshan, Maoyang), Yunnan province (Mengla), Liaoning province (Xishan), and Taiwan (Taipei, Kaohsiung) (Figure 1).

2.2. Mosquito DNA Preparation and Amplification. Nucleic acid was extracted using the TIANamp Genomic DNA Kit after removing the head of mosquitos. We then used a pair of universal primers including LCO1490 (5′-GGTCAACAAATCATAAAGATATTGG-3′) and HCO2198 (5′-TAAACTTCAGGGTGACCAAAAAAT-CA-3′) to amplify them [17]. The reaction system was set as follows: Ex Taq TaKaRa 0.25μL, 10XEx Taq Buffer 5μL, dNTP Mixture 4μL, Template DNA 3μL (if the light belt can be properly increased), and upper and lower primer (10 μM) with each 1μL, making a total volume of 50μL by adding ddH_2O. In addition, the amplification conditions are set as follows: predenaturation at 94°C for 3 min, denaturation at 94°C for 30 s, annealing at 55°C for 30 s, extension at 72°C for 45 sec, a total of 30 cycles, and final elongation at 72°C for 5 min.

2.3. PCR Product Purification and Cloning. The PCR products were cloned by Universal DNA Purification kit, equipped with LB liquid medium, LB solid medium, and 50 mg/ml IPTG. The PCR products were digested with PCR Identification Kit for Recombinant pGM-T Clone, which were sent out for sequencing.

2.4. Data Analysis

Sequence Verification and Alignment. By referring to the NCBI sequence database, homology search was performed against existing COI gene sequences of *Aedes albopictus* in GenBank, and sequences with identity greater than or equal to 98% were selected [18].

Sequence Analysis. The aligned sequences were input into the multiple sequence alignment tool Clustal X (v1.8) for multiple sequence alignment [19, 20]. The MEGA6 software package was used to summarize the statistics of sequence characteristics, including base content, the number of transversions, and transitions, and calculate the genetic distance of the sequences [21]. Taking *Chironomidae nepeanensis* (GenBank: KC750313.1) as the outgroup, we constructed Neighbor-Joining (NJ) tree and Maximum Likelihood (ML) tree according to the genetic distance Kimura 2-parameter method and bootstrapped 1000 times to test the reliability of the branch trees [21–23]. The DnaSP 5.0 software [24] was used to identify the polymorphism sites and calculate the haplotype diversity and nucleotide diversity of each

Table 1: Haplotype diversity and nucleotide diversity of different geographical strains.

Collecting places	Number of samples	Number of haplotypes	Haplotype type (n)	Polymorphism sites (s)	Haplotype diversity (Hd)	Nucleotide diversity (Pi)
Fujian	20	12	h1(2),h2∗(7),h3(1),h4(2),h5(1), h6(1),h7(1),h8(1),h9(1),h10(1), h11(1),h12(1)	66	0.87895	0.03316
Guangdong	20	7	h2∗(13),h6∗(1),h13(1),h14(1), h15(1),h16(1)h17(2)	8	0.58421	0.00173
Hainan	19	9	h18∗(5),h19∗(5),h20(1),h21(1), h22(3),h23(1),h24(1),h25(1), h26(1),	11	0.86550	0.00265
Yunnan	20	11	h18∗(4),h19∗(1),h27(1),h28(6), h29(1),h30(1),h31(1),h32(1), h33(1),h34(2),h35(1)	13	0.88421	0.00372
Taiwan	22	8	h2∗(9),h36(1),h37(7),h38(1), h39(1),h40(1),h41(1),h42(1)	11	0.75325	0.00215
Liaoning	5	1	h2∗(5)	0	0.00000	0.00000

Note. ∗ denotes shared haplotypes.

geographical population of *Aedes albopictus*. The median joining method of Network4.6 software was used to construct haplotype pedigree network diagram. The analysis of molecular variance (AMOVA v3.1) was used to calculate the genetic differentiation within and between populations, including Fst (F-statistics) and Nm (Nm = (L − Fst)/4Fst) [25]. Arlequin was used to do mismatch analysis and neutrality test. We then calculated sum of squared deviation (SSD) and Harpending's Raggedness (HR) indices through mismatch analysis and constructed observation simulation model [26, 27], drew bifurcation point distribution, and explored the evolutionary history of *Aedes albopictus* population. Using the Tajima's D [28] and Fu's Fs values [29] of the neutrality test, we further explored the population expansion mechanism.

The ancestral population size (θ) is calculated as follows: $\theta = 2Nu$, where N represents the effective number of female mosquitos in the population, u represents the mutation rate per generation, and $u = \mu k$ (μ is the mutation rate per point per generation and k is the base number of the analyzed sequences). We used $\tau = 2ut$ to estimate the approximate generation number (t) of population expansion [26] and set the mutation rate per point per generation of *Aedes albopictus* which is 1×10^{-8} [30].

Mantel test was performed by online software IBD Web Service (IBDWS) [31] to evaluate the correlation between genetic differentiation coefficient and geographical distance.

3. Results

3.1. Sequence Features. The mitochondrial COI gene was obtained with length 709bp. According to the sequence alignment of *Aedes albopictus* in NCBI database, the sequence identity was 99%-100% showing very small difference among the 106 sequences.

After sequence alignment, we took the 658 bp fragment (with the universal primer removed) for subsequent analysis. The overall base composition of the fragment is A (28.5%), T (39.2%), C (16.7%), and G (15.6%). The A+T content (67.7%) is greater than that of G+C (32.3%). There were 90 (13.68%) polymorphism sites, of which 57 (63.33%) are transitions, 28 (31.11%) are transversions, and 5 (5.56%) are hypervariable loci (Table S1).

3.2. The Relationships among Haplotypes of Different Geographical Strains. We obtained 42 haplotypes on 6 geographical strains of *Aedes albopictus MT-COI* gene (Table 1), 4 (9.52%) of which are shared haplotypes among different populations. They are h2, h6, h18, and h19, respectively: h2 is shared by Fujian, Guangdong, Taiwan, and Liaoning, h6 is shared by Fujian and Guangdong, and h18 and h19 are shared by Hainan and Yunnan. The result indicates that there are gene exchanges among the populations. However, it can be seen from the exclusive haplotypes that *Aedes albopictus* has some genetic differentiation. Haplotype diversity (Hd) and nucleotide diversity (Pi) were used to indicate the degree of haplotype differentiation and the degree of nucleotide sequence variation. The haplotype diversity of *Aedes albopictus* is 0.882 and nucleotide diversity is 0.01017. The haplotype diversity of geographical strains is ordered as Yunnan > Fujian > Hainan > Taiwan > Guangdong > Liaoning strains. The nucleotide diversity is ordered as Fujian > Yunnan > Hainan > Taiwan > Guangdong > Liaoning strains. Fujian strains had the largest number of polymorphism sites (s = 66) and haplotype number (n = 12) and thus also had the largest nucleotide diversity (Pi = 0.03316). Its haplotype diversity is 0.87895, only slightly smaller than that of Yunnan strains (Hd = 0.88421). There were no differences between the 5 individuals in Liaoning populations.

The median joining method in software Network4.6 was used to construct haplotypes pedigree network of 42 haplotypes from 6 geographical strains of *Aedes albopictus* mtDNA-COI gene (Figure 2). In the network, each circle represents a haplotype, the size of the circle represents the number of homozygous haplotypes, different colors

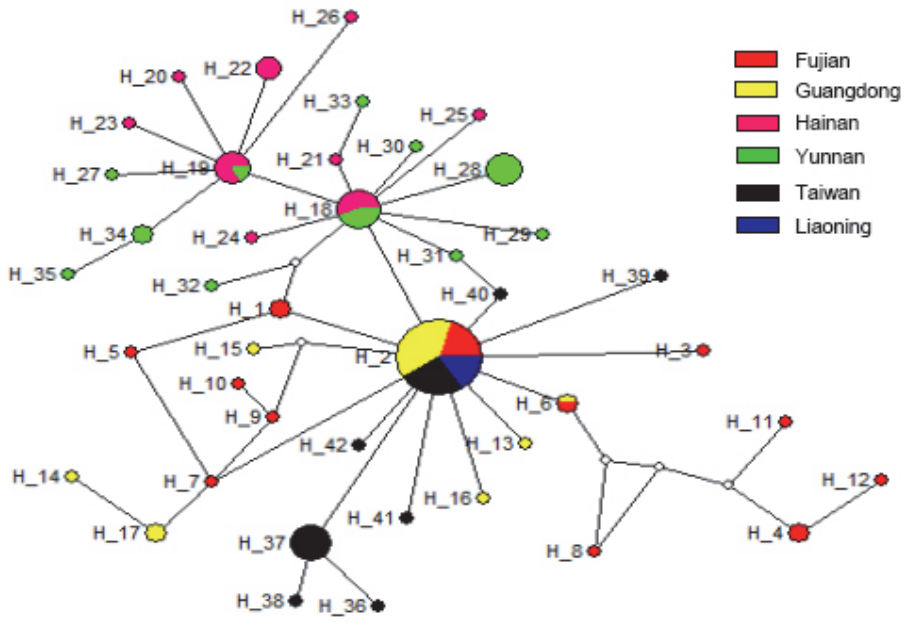


Figure 2: The haplotypes pedigree network diagram of different geographical strains of *Aedes albopictus MT-COI* gene.

Table 2: The genetic distances within and between different geographical strains.

Collecting places	The genetic distances within populations	The genetic distances between populations				
		Fujian	Guangdong	Hainan	Yunnan	Taiwan
Fujian	0.03535					
Guangdong	0.00173	0.02187				
Hainan	0.00266	0.02403	0.00421			
Yunnan	0.00374	0.02473	0.00466	0.00361		
Taiwan	0.00216	0.02252	0.00223	0.00462	0.00504	
Liaoning	0.00000	0.02122	0.00091	0.00337	0.00374	0.00132

indicate different geographical haplotypes, and the blank circles indicate undetected haplotypes. As can be seen from the network, there is a certain level of parallel evolution between the haplotypes, suggesting the expansion of *Aedes albopictus* in history. Among them, unchecked haplotypes are rare (5), but shared haplotypes H2 (34, 32.08%), H18 (9, 8.49%), and H19 (6, 5.66%) are highly distributed in the population, which may be the sources of expansion. It can be seen that Yunnan + Hainan strains and other geographical strains are roughly divided into two categories, but there are individual haplotypes intersecting with each other.

MEGA6 was used to calculate the genetic distances among *MT-COI* gene haplotypes of different geographical strains and the NJ and ML methods were used to construct the phylogenetic trees respectively with 1,000 times bootstrapping to assess branch confidence. The trees constructed from the two methods are quite similar, so we took NJ tree for illustration (Figure 3). As can be seen, part of Fujian haplotypes was clustered into one branch (H4, H11, and H12) with high confidence and the rest (Hainan + Yunnan) (Guangdong + Taiwan + Liaoning + part of Fujian) were clustered into their respective branches with low confidence. The NJ tree roughly coincides with the haplotypes pedigree network, suggesting a certain correlation between haplotype clusters and geographical distribution. Further experiment should be done for Liaoning strains due to the small sample size and number of single haplotype.

3.3. Population Genetic Structure. The MEGA6 software package was used to calculate the genetic distances of different geographical strains (Table 2). Genetic distance is one of the indicators to measure the genetic diversity within a population. It can be seen from the table that the genetic distance within each geographical population is ordered as Fujian > Yunnan > Hainan > Taiwan > Guangdong > Liaoning. The genetic distance between populations ranges from 0.00091 to 0.02473 with an average of 0.01925. The genetic distance between Fujian and Yunnan strains is largest. The genetic distances between Fujian and other geographical strains are more than 0.02, indicating the significant difference between Fujian and other regions. The genetic distance between Liaoning and Guangdong strains is the smallest.

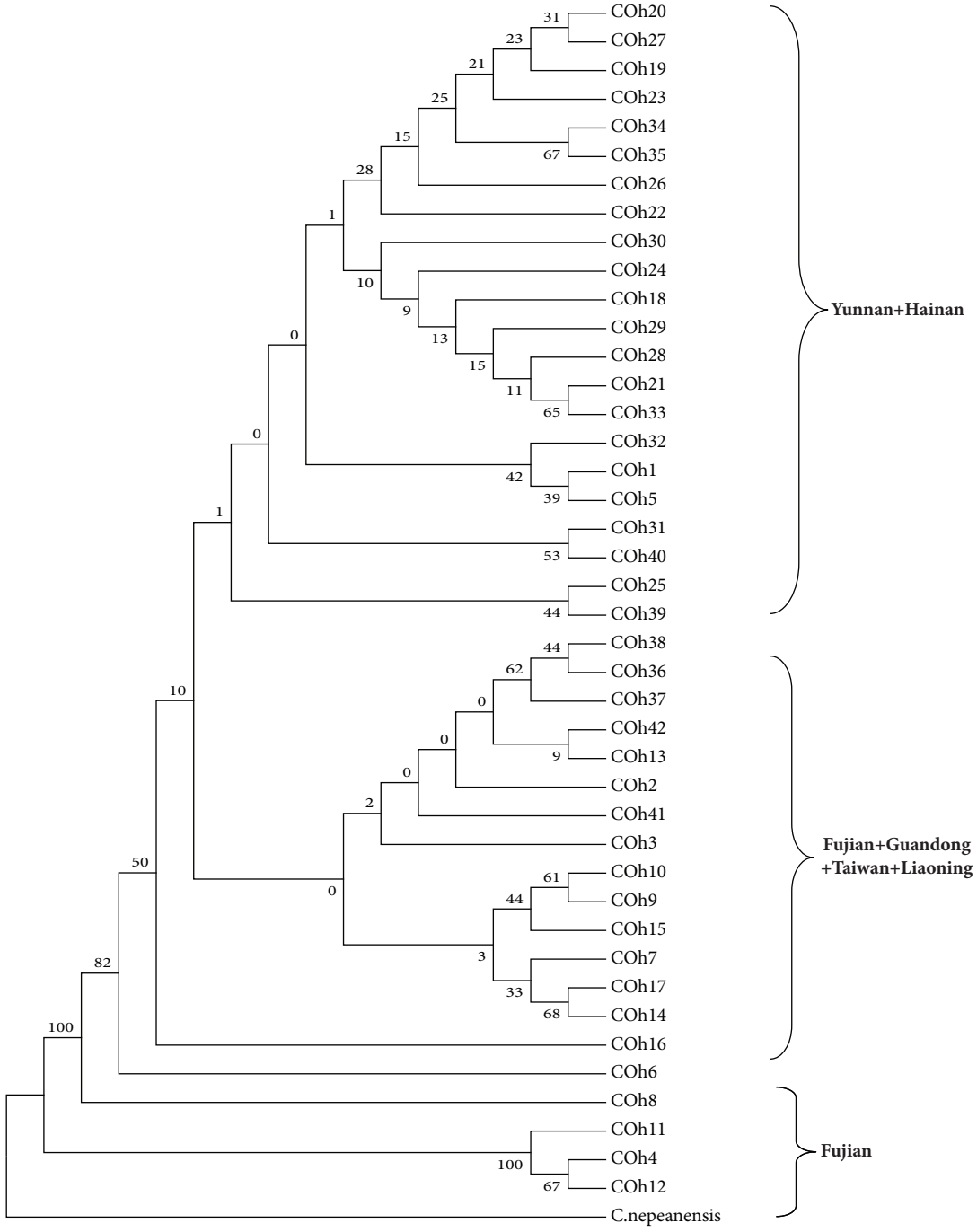


Figure 3: The NJ tree of *Aedes albopictus MT-COI* gene haplotypes from different geographical strains.

Arlequin 3.1 was used to calculate the genetic differentiation coefficients (Fst) among geographical strains and the results were summarized in Table 3. Fst is an indicator to measure the genetic differentiation between populations. Great Fst value means greater genetic differentiation and less gene exchange. When Nm is greater than 1, gene exchange can prevent genetic differentiation between populations caused by genetic drift. The Nm values between Fujian, Guangdong, and Taiwan strains are all greater than 1, indicating frequent genetic exchanges between these geographical populations. The Fst value (0.11363) and Nm value (1.95012) between Hainan and Yunnan strains indicate the smallest genetic differentiation and frequent gene exchange between them. However, the Nm values of the two geographical strains with other geographical strains are less than 1, indicating limited genetic exchanges.

The AMOVA analysis (Table 4) shows that the variation within populations (78.13%) is greater than that between populations (21.87%), which suggests that the genetic differentiation of *Aedes albopictus* population structure mainly

Table 3: The Fst (lower triangle) and Nm (upper triangle) values of different geographical strains of *Aedes albopictus MT-COI* gene.

Collecting places	Fst\Nm					
	Fujian	Guangdong	Hainan	Yunnan	Taiwan	Liaoning
Fujian		1.40596	0.96993	0.94554	1.17849	9.35430
Guangdong	0.15097		0.27134	0.35618	1.71757	−4.09084
Hainan	0.20493	0.47953		1.95012	0.27040	0.29583
Yunnan	0.20911	0.41242	0.11363		0.34867	0.48844
Taiwan	0.17501	0.12706	0.48040	0.41759		7.06422
Liaoning	0.02603*	−0.06509*	0.45802	0.33855	0.03418*	

Note. * represents p>0.05.

Table 4: The AMOVA analysis of different geographical strains of *Aedes albopictus MT-COI* gene.

Source of variation	Degree of freedom	Variation components	Percentage of variation (%)
Between populations	5	0.76116	21.87
Within population	100	2.71898	78.13

Note. Fst: 0.21871, Nm: 0.893.

Table 5: The mismatch analysis of different geographical strains of *Aedes albopictus*.

	Overall	Fujian	Guangdong	Hainan	Yunnan	Taiwan	Liaoning
SSD	0.520	0.00*	0.00*	0.50	0.20	0.45	0.00*
HR	0.790	1.00	1.00	0.25	0.30	0.40	0.00*

Note. * represents P<0.05.

comes from the population interior. The total of Fst value (0.21871) and Nm value (0.893) indicates that the overall gene flow of *Aedes albopictus* failed to prevent the population differentiation caused by genetic drift and there is a certain level of genetic differentiation in the population.

3.4. Group Dynamics. We used the mismatch analysis of Arlequin (Table 5) to assess the reliability of expansion through SSD and HR parameters. If the difference between the two parameters is not statistically significant (i.e., P > 0.05), the assumption of population expansion could not be rejected, which is in line with the original group expansion hypothesis. Overall analysis from *Aedes albopictus* indicates that the p-values of SSD and HR are all greater than 0.05, suggesting an expansion of *Aedes albopictus* in history. From the perspective of geographical strains, the p-values of SSD and HR of Hainan, Yunnan, and Taiwan strains are all greater than 0.05, suggesting an expansion of *Aedes albopictus* in these areas in history. In Fujian and Guangdong, only p-value of HR parameter is greater than 0.05, suggesting that *Aedes albopictus* in these areas do not have significant expansion in history.

We mapped the bifurcation distribution of different geographical strains of *Aedes albopictus MT-COI* gene (Figure 4) and observed a fitting between the expected and the observed values, which also indicates population expansion. Generally speaking, the population is suggested to be in balance when the observed values of bifurcation distribution do not coincide with the expected ones (i.e., the figure will show multipeaks); otherwise (i.e., single peak) it will indicate a population expansion [32]. Specifically, the observed values of Hainan, Yunnan, and Taiwan strains coincide with the expected values and showed a single peak distribution, but the distributions of Fujian and Guangdong strains are not consistent. The overall bifurcation distribution of *Aedes albopictus* presents a single peak distribution, indicating population expansion.

Tajima's D value and Fu's Fs value (Table 6) were calculated using the neutrality test of the Arlequin software (Table 6). In theory, negative Tajima'D and Fu's Fs values will indicate a population expansion in history. In our case, the *D* value (−1.991) and Fs value (−24.983) of overall population both reach a significant level, indicating a significant population expansion in history. As for specific geographical strains, except for Fujian strains with positive D value and Yunnan strains without significant negative D value, other geographical strains are with statistically significant negative D and Fs values. The general results are consistent with the mismatch analysis.

According to the formula $\theta = 2Nu$, the effective size of female mosquito population was estimated based on θ_0 and θ_1 before and after expansion. When $\theta_0 = 0.443$ and $\theta_1 = 99999$, the effective number of female mosquitos in the population is estimated to be $3.4 \times 10^4 \sim 7.6 \times 10^9$. According to τ (95% CI) = 2ut = 3.943, the population expansion occurred in 3.0×10^5 years ago.

3.5. The Relationship between Genetic Differentiation and Geographical Distance. The Mantel test was performed on the geographical distances between different collection points and *Aedes albopictus MT-COI* gene sequences. The correlation curve between genetic differentiation coefficients Fst and

Table 6: The neutrality test of different geographical strains.

Collecting places	Tajima's *D*		Fu's *Fs*		Tau (95%)
	D value	P value	*FS* value	P value	
Overall	−1.991	0.002	−24.983	0.000	3.943
Fujian	0.703	0.280*	−6.878	0.003	1.008
Guangdong	−1.672	0.037	−29.191	0.000	0.000
Hainan	−1.604	0.038	−27.472	0.000	2.242
Yunnan	−1.205	0.119*	−26.629	0.000	3.820
Taiwan	−1.837	0.019	−27.981	0.000	2.070
Liaoning	0.000	-	∞	-	0.000

Note: *P>0.05.

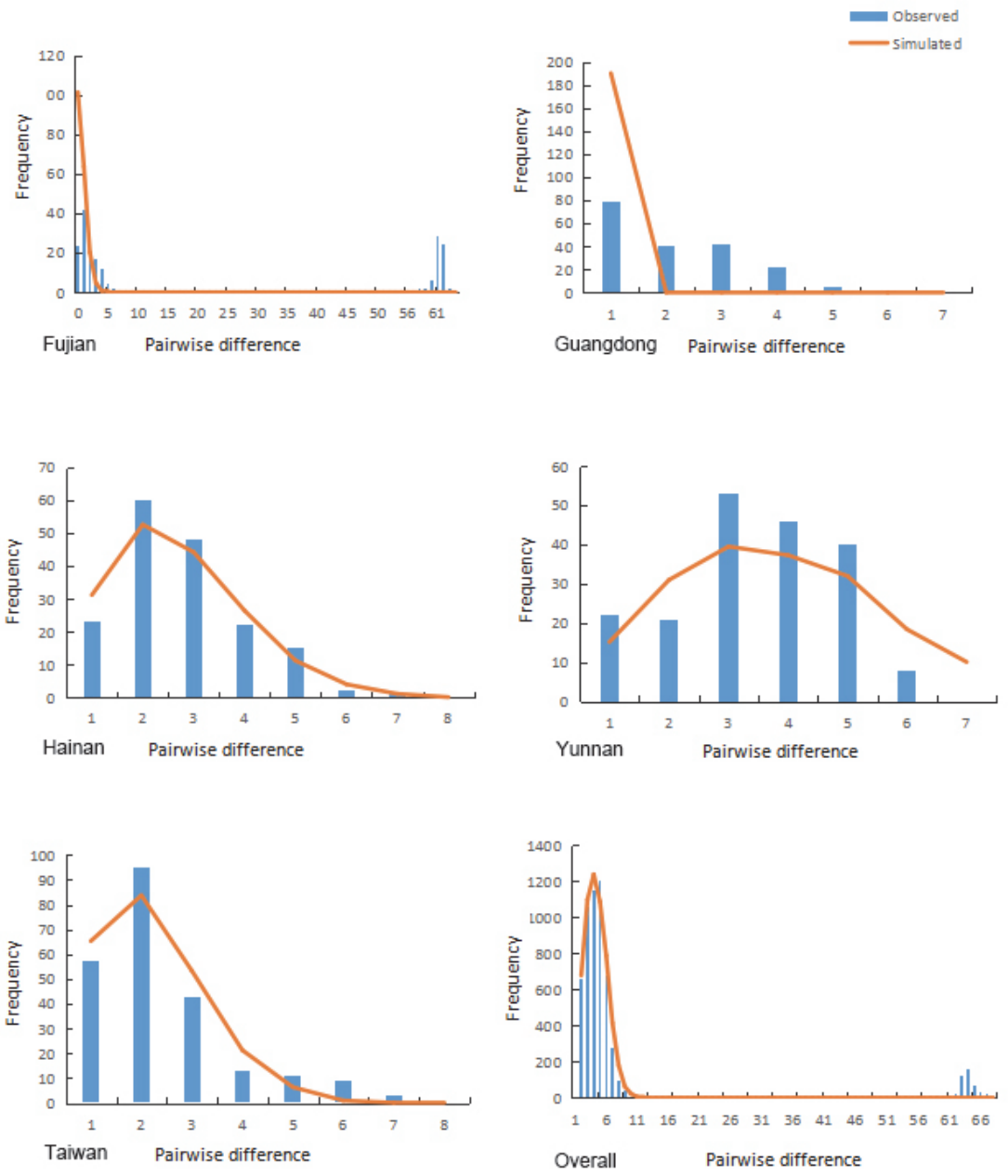


Figure 4: Mismatch distribution of *Aedes albopictus MT-COI* gene from different geographical regions.

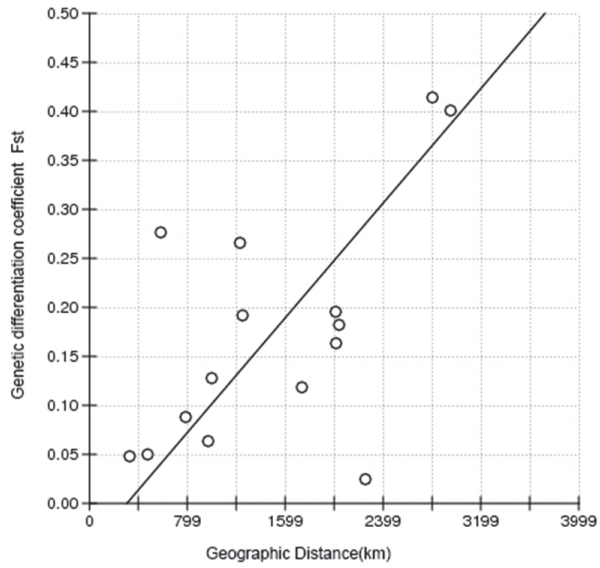


Figure 5: The relationship between genetic differentiation of *Aedes albopictus MT-COI* gene and geographical distance.

geographical distances was illustrated in Figure 5. The correlation between population Fst and geographical distances (r = 0.5789, p = 0.0120) was statistically significant, indicating a positive correlation between genetic differentiation and geographical distance.

4. Discussion

The genetic diversity of insect species can arise from two sources: internal and external causes. The internal causes are genetic mutation and adaptability of insect itself while the external factors are closely related to ecological environment of insects. The accumulation of genetic difference will lead to reproductive isolation and thus the formation of new species. A few scholars have studied the diversity of *Aedes albopictus* by different methods. Preliminary studies have confirmed that there is a certain degree of genetic differentiation among *Aedes albopictus* populations. Therefore, further studies on genetic variation of *Aedes albopictus* population are needed to provide valuable information for the prevention and control of dengue fever. Also, new mosquito species can become the vector of new diseases, and it is critical to surveil the mosquito species in a geographical region.

Haplotype diversity (Hd) and nucleotide diversity (Pi) are two important indicators to measure the diversity of a species population [33]. The haplotype diversity and nucleotide diversity of *Aedes albopictus* in this study (Hd = 0.882, Pi = 0.01017) are higher than those in *Aedes aegypti* (Hd = 0.740 ± 0.017, P = 0.0065 ± 0.0003), indicating high diversity *of Aedes albopictus* population. *Aedes albopictus* exhibited high Hd and low Pi mode probably because of rapid population expansion after the bottleneck effect. The formation of a new haplotype is due to a single base variation, which has less effect on nucleotide diversity, so increasing nucleotide diversity requires more time to accumulate than increasing haplotype diversity. The bottleneck effect can eliminate the accumulation of nucleotide diversity in the past but can also cause mutations at the base site, resulting in a pattern of high Hd and low Pi. The possible reasons for the bottleneck of *Aedes albopictus* may be as follows: on the one hand, in terms of neutral test of *Aedes albopictus* (D = −1.99, P = 0.002; Fs = −24.98, P = 0.000) and the bifurcation distribution representing a single peak, *Aedes albopictus* has been or is in a state of population expansion. On the other hand, dengue fever has caused great harm to human health since it was recognized that controlling the number of media mosquitos is critical for controlling the spread of the disease. Therefore, a wide range of remediation media mosquitos have a greater impact on *Aedes albopictus* population.

In this study, the genetic distances between different geographical populations ranged from 0.00091 to 0.02473, which are higher than those of *Aedes albopictus* in different regions of Guangzhou about 0.000~0.007 in a previous study. Among them the genetic distances between Fujian and other geographical strains were greater than 0.02 (Table 2), which is greater than the difference of less than 2% of the genetic distance within 98% of the species proposed by Hebert [10], indicating a large genetic difference between Fujian *Aedes albopictus* and other geographical strains. Haplotype diversity, nucleotide diversity, and genetic distance are all indicators to reflect the genetic diversity of population. The results of the three indicators in this study (Tables 1 and 2) basically showed the largest value of Fujian strains, indicating Fujian population with the largest genetic diversity. In the pedigree network diagram, Haplotypes h2 is found to be a shared haplotype among Fujian, Guangdong, Taiwan, and Liaoning and developed into other haplotypes through direct or indirect evolution of 1~5 steps. However, the NJ tree shows that 3 haplotypes (h4, h11, and h12) of Fujian clustered independently into one category with high confidence, which further indicates the result that Fujian strains of *Aedes albopictus* have obvious genetic structure changes and also verified early studies. Those studies suggested that there is a certain degree of genetic differentiation of different geographical strains of *Aedes albopictus* in Fujian province. Due to the farthest flight distance of *Aedes albopictus* not exceeding 500 meters, it is very difficult to invade from the exterior in short time. The reason for this phenomenon may be related to the ecological and climatic environment of *Aedes albopictus* in Fujian. Fujian is located in a subtropical maritime monsoon climate, whose complex topography forms a variety of local climate, thus forming different ecological environments, which provide favorable conditions for breeding and living *Aedes albopictus*. Fujian geographical strains of *Aedes albopictus* haplotypes diversity, the causes of specific haplotype, and whether there is a special significance in the mosquito-borne disease transmission need further study.

Morton [34] believed that when the gene flow Nm value is greater than 1, it means that the gene exchange is frequent, which can prevent the interpopulation differentiation caused by genetic drift. When the value is less than 1, it indicates that gene exchange is blocked. In this study, the *Aedes albopictus* Nm value is less than 1 (Nm = 0.893), indicating

that the level of gene exchange failed to prevent population differentiation caused by genetic drift and there is a certain level of genetic differentiation between populations. In terms of Nm value of different geographical strains (Table 3), the Nm value between Hainan and Yunnan strains is greater than 1, indicating that *Aedes albopictus* in the two places had frequent genetic exchange. However, the Nm values between (Yunnan + Hainan strains) and Guangdong, Taiwan and Liaoning strains are less than 1, indicating that *Aedes albopictus* gene exchanges between (Yunnan + Hainan) and the four regions are hindered. This is consistent with NJ tree that Yunnan + Hainan was clustered into one category. The NJ tree and gene flow show that the genetic exchanges frequently happen among Taiwan, Fujian, and Guangdong mosquitos, suggesting that Taiwan strains have the same type of geographical populations as Fujian and Guangdong strains. No relevant reports have been found yet.

Finally, this study shows that the genetic differences within *Aedes albopictus* population are larger than those between populations. Due to the different ecological and climatic conditions in different regions, there are various kinds of natural barriers. However, *Aedes albopictus* belongs to the semihabitat mosquito and is closely related to human activities, the geographical barrier failed to completely prevent the gene flow, so genetic differentiation mainly which comes from internal and genetic differentiation among the populations is not high. The result of the Mantel test ($r = 0.5789$, $p = 0.0120$) shows that the degree of CO I gene sequence variation is positively correlated with geographical distance. Further experiments will be conducted to confirm this conclusion. Giving more and more genetic sequences of mosquitos in different regions has been published; it would be interesting to study the genetic differences of mosquito populations throughout China or across a continent in the future.

5. Conclusion

In conclusion, our results suggest that (1) there are genetic differences among *Aedes albopictus* populations in different geographical regions; (2) there is a positive correlation between the genetic differences of *Aedes albopictus* population and their geographical distances; and (3) DNA barcoding of *MT-COI* gene is an effective method to study the genetic structure of *Aedes albopictus*.

Conflicts of Interest

The authors have declared no conflicts of interest.

Authors' Contributions

Bo Gao conceived and designed the experiments. Yiliang Fang and Qianqian Qian performed the experiments. Yiliang Fang analyzed the data. Yiliang Fang and Bo Gao wrote the paper. Jianqing Zhang and Baohai Xue contributed to the discussion and revision of the paper. All authors have approved the final manuscript.

Supplementary Materials

Table S1: variable sites in *MT-COI* haplotype sequences of *Aedes albopictus* in different geographical regions. The dot "." denotes consensus nucleotide with coI h1. *(Supplementary Materials)*

References

[1] N. G. Gratz, "Critical review of the vector status of *Aedes albopictus*," *Medical and Veterinary Entomology*, vol. 18, no. 3, pp. 215–227, 2004.

[2] T. Coffinet, J. R. Mourou, B. Pradines et al., "First record of Aedes albopictus in Gabon," *Journal of the American Mosquito Control Association*, vol. 23, no. 4, pp. 471-472, 2007.

[3] D. A. Yee, S. A. Juliano, and S. M. Vamosi, "Seasonal photoperiods alter developmental time and mass of an invasive mosquito, *Aedes albopictus* (*Diptera*: *Culicidae*), across its north-south range in the United States," *Journal of Medical Entomology*, vol. 49, no. 4, pp. 825–832, 2012.

[4] L. A. Ganushkina, I. V. Patraman, G. Rezza, L. Migliorini, S. K. Litvinov, and V. P. Sergiev, "Detection of *Aedes aegypti, Aedes albopictus, and Aedes koreicus* in the Area of Sochi, Russia," *Vector-Borne and Zoonotic Diseases*, vol. 16, no. 1, pp. 58–60, 2016.

[5] C. E. Smith, "The history of dengue in tropical Asia and its probable relationship to the mosquito *Aedes aegypti*," *The Journal of tropical medicine and hygiene*, vol. 59, no. 10, pp. 243–251, 1956.

[6] P. Reiter, "*Aedes albopictus* and the world trade in used tires, 1988-1995: The shape of things to come?" *Journal of the American Mosquito Control Association*, vol. 14, no. 1, pp. 83–94, 1998.

[7] X.-X. Guo, C.-X. Li, Y.-M. Zhang et al., "Vector competence of *Aedes albopictus* and *Aedes aegypti* (*Diptera*: *Culicidae*) for the DEN2-FJ10 and DEN2-FJ11 strains of the dengue 2 virus in Fujian, China," *Acta Tropica*, vol. 161, pp. 86–90, 2016.

[8] P. D. N. Hebert, A. Cywinska, S. L. Ball, and J. R. DeWaard, "Biological identifications through DNA barcodes," *Proceedings of the Royal Society B Biological Science*, vol. 270, no. 1512, pp. 313–321, 2003.

[9] F. O. Costa, J. R. DeWaard, J. Boutillier et al., "Biological identifications through DNA barcodes: The case of the Crustacea," *Canadian Journal of Fisheries and Aquatic Sciences*, vol. 64, no. 2, pp. 272–295, 2007.

[10] P. D. N. Hebert, S. Ratnasingham, and J. R. DeWaard, "Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species," *Proceedings of the Royal Society B Biological Science*, vol. 270, supplement 1, pp. S96–S99, 2003.

[11] R. D. Ward, T. S. Zemlak, B. H. Innes, P. R. Last, and P. D. N. Hebert, "DNA barcoding Australia's fish species," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 360, no. 1462, pp. 1847–1857, 2005.

[12] A. Cywinska, F. F. Hunter, and P. D. N. Hebert, "Identifying Canadian mosquito species through DNA barcodes," *Medical and Veterinary Entomology*, vol. 20, no. 4, pp. 413–424, 2006.

[13] E. L. Clare, B. K. Lim, M. D. Engstrom, J. L. Eger, and P. D. Hebert, "DNA barcoding of Neotropical bats: species identification and discovery within Guyana," *Molecular Ecology Resources*, vol. 7, no. 2, pp. 184–190, 2007.

[14] K. C. R. Kerr, M. Y. Stoeckle, C. J. Dove, L. A. Weigt, C. M. Francis, and P. D. N. Hebert, "Comprehensive DNA

barcode coverage of North American birds," *Molecular Ecology Resources*, vol. 7, no. 4, pp. 535–543, 2007.

[15] N. P. Kumar, A. R. Rajavel, R. Natarajan, and P. Jambulingam, "DNA barcodes can distinguish species of indian mosquitoes (*Diptera*: *Culicidae*)," *Journal of Medical Entomology*, vol. 44, no. 1, pp. 1–7, 2007.

[16] C. G. Moore and C. J. Mitchell, "*Aedes albopictus* in the United States: ten-year presence and public health implications," *Emerging Infectious Diseases*, vol. 3, no. 3, pp. 329–334, 1997.

[17] O. Folmer, M. Black, W. Hoeh, R. Lutz, and R. Vrijenhoek, "DNA primers for amplification of mitochondrial *cytochrome c oxidase* subunit I from diverse metazoan invertebrates," *Molecular Marine Biology and Biotechnology*, vol. 3, no. 5, pp. 294–299, 1994.

[18] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler, "GenBank," *Nucleic Acids Research*, vol. 33, pp. D34–D38, 2005.

[19] R. Chenna, H. Sugawara, T. Koike et al., "Multiple sequence alignment with the Clustal series of programs," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3497–3500, 2003.

[20] J. Yang and L. Zhang, "Run probabilities of seed-like patterns and identifying good transition seeds," *Journal of Computational Biology*, vol. 15, no. 10, pp. 1295–1313, 2008.

[21] S. Kumar, G. Stecher, and K. Tamura, "MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets," *Molecular Biology and Evolution*, vol. 33, no. 7, pp. 1870–1874, 2016.

[22] J. Yang, S. Grünewald, and X.-F. Wan, "Quartet-net: A quartet-based method to reconstruct phylogenetic networks," *Molecular Biology and Evolution*, vol. 30, no. 5, pp. 1206–1217, 2013.

[23] J. Yang, S. Grünewald, Y. Xu, and X.-F. Wan, "Quartet-based methods to reconstruct phylogenetic networks," *BMC Systems Biology*, vol. 8, article 21, 2014.

[24] J. Rozas, J. C. Sánchez-DelBarrio, X. Messeguer, and R. Rozas, "DNAsp, DNA polymorphism analyses by the coalescent and other methods," *Bioinformatics*, vol. 19, no. 18, pp. 2496-2497, 2003.

[25] L. Excoffier, P. E. Smouse, and J. M. Quattro, "Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data," *Genetics*, vol. 131, no. 2, pp. 479–491, 1992.

[26] A. R. Rogers and H. Harpending, "Population growth makes waves in the distribution of pairwise genetic differences," *Molecular Biology and Evolution*, vol. 9, no. 3, pp. 552–569, 1992.

[27] H. C. Harpending, "Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution," *Human Biology*, vol. 66, no. 4, pp. 591–600, 1994.

[28] F. Tajima, "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism," *Genetics*, vol. 123, no. 3, pp. 585–595, 1989.

[29] Y.-X. Fu, "Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection," *Genetics*, vol. 147, no. 2, pp. 915–925, 1997.

[30] J. R. Powell, A. Caccone, G. D. Amato, and C. Yoon, "Rates of nucleotide substitution in Drosophila mitochondrial DNA and nuclear DNA are similar," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 83, no. 23, pp. 9090–9093, 1986.

[31] J. L. Jensen, A. J. Bohonak, and S. T. Kelley, "Isolation by distance, web service," *BMC Genetics*, vol. 6, 2005.

[32] M. Slatkin and R. R. Hudson, "Pairwise comparisons of mitochondrial DNA sequences in stable & exponentially growing populations," *Genetics*, vol. 129, no. 2, pp. 555–562, 1991.

[33] R. C. Vrijenhoek, "Genetic diversity and fitness in small populations," *Conservation Genetics*, pp. 37–53, 1994.

[34] N. E. Morton, "Isolation by distance in human populations," *Annals of Human Genetics*, vol. 40, no. 3, pp. 361–365, 1977.

Hindawi
BioMed Research International
Volume 2018, Article ID 1096079, 10 pages
https://doi.org/10.1155/2018/1096079

Research Article

Isolation of a Reassortant H1N2 Swine Flu Strain of Type "Swine-Human-Avian" and Its Genetic Variability Analysis

Long-Bai Wang, Qiu-Yong Chen, Xue-Min Wu, Yong-Liang Che, Cheng-Yan Wang, Ru-Jing Chen, and Lun-Jiang Zhou

Institute of Animal Husbandry and Veterinary Medicine, Fujian Academy of Agriculture Sciences, Fujian Animal Disease Control Technology Development Center, Fuzhou, Fujian 350013, China

Correspondence should be addressed to Lun-Jiang Zhou; lunjiang@163.com

Received 3 January 2018; Accepted 26 February 2018; Published 29 May 2018

Academic Editor: Jialiang Yang

We isolated an influenza strain named A/Swine/Fujian/F1/2010 (H1N2) from a pig suspected to be infected with swine flu. The results of electron microscopy, hemagglutination (HA) assay, hemagglutination inhibition (HI) assay, and whole genome sequencing analysis suggest that it was a reassortant virus of swine (H1N1 subtype), human (H3N2 subtype), and avian influenza viruses. To further study the genetic evolution of A/Swine/Fujian/F1/2010 (H1N2), we cloned its whole genome fragments using RT-PCR and performed phylogenetic analysis on the eight genes. As a result, the nucleotide sequences of HA, NA, PB1, PA, PB2, NP, M, and NS gene are similar to those of A/Swine/Shanghai/1/2007(H1N2) with identity of 98.9%, 98.9%, 99.0%, 98.6%, 99.0%, 98.9%, 99.3%, and 99.3%, respectively. Similar to A/Swine/Shanghai/1/2007(H1N2), we inferred that the HA, NP, M, and NS gene fragments of A/Swine/Fujian/F1/2010 (H1N2) strain were derived from classical swine influenza H3N2 subtype, NA and PB1 were derived from human swine influenza H3N2 subtype, and PB2 and PA genes were derived from avian influenza virus. This further validates the role of swine as a "mixer" for influenza viruses.

1. Introduction

Swine influenza is an acute and highly contagious flu caused by swine influenza virus (SIV), which infects human and pigs. Its clinical symptoms in swine include high fever, decreased appetite, lethargy, sneezing, coughing, and difficulty in breath [1]. In virus classification, swine influenza viruses are single-stranded negative-sense RNA viruses belonging to the family Orthomyxoviridae. The genome of swine influenza virus is composed of 8 gene fragments with different sizes, namely, PB2, PB1, PA, HA, NP, NA, M, and NS, respectively [2]. The 8 genomic segments of different influenza viruses can be randomly shuffled to generate new reassortant viruses [3]. As pigs possess both NeuAc-2,3Gal and NeuAc-2,6Gal receptors, they can infect both human and avian influenza viruses [3]. As a result, pigs become a "mixer" for reassortment of influenza viruses and an "incubator" for emergence of new influenza strains [4, 5]. Pigs also play an important role in the ecological distribution and genetic evolution of influenza viruses. For example, the Sydney-like H3N2 mutant isolated in the United States in 1998 is an "avian-swine-human" reassortant strain [6]. In addition, the pandemic H1N1 influenza virus originated in April 2009 is a "swine-human-avian" reassortant virus [7]. In fact, two out of the four most serious influenza pandemics in the history are caused by reassortment of gene segments from avian influenza viruses [8, 9].

At present, there are three subtypes of swine influenza in the world, namely, H1N1, H3N2, and H1N2, which can be further divided into classic swine H1N1, avian H1N1, human-like H3N2, reassortant H3N2, and reassortant H1N2 [10, 11]. In addition to these three subtypes, strains of subtypes H5N1, H9N2, and H3N1 have also been isolated in swine populations; however they are not prevalent enough in swine populations to establish stable lineages [12–14]. The H3N2 swine influenza subtype was first reported in Taiwan in 1969, and the H3N2 swine influenza viruses were isolated from pigs in Hong Kong in 1970 [15]. After that, they have become widespread in swine worldwide. H1N2 influenza subtype was first isolated from Japanese swine herds in 1978 and

subsequently widespread in other countries such as France, UK, and USA. The H1N2 swine influenza viruses are diverse in gene origin, which results in their different levels of prevalence and danger. In China, H1N2 swine influenza virus was first isolated in Zhejiang in 2004 and then isolated in Guangxi, Shanghai, Guangdong, and other places from 2005 to 2010.

From 2009 to 2013, our laboratory carried out epidemiological study of swine flu in Fujian Province. An influenza strain was isolated from 13 pigs suspected to be infected. We then performed electron microscopy, hemagglutination (HA) assay, hemagglutination inhibition (HI) assay, and whole genome sequencing analysis to infer its biological characteristics and evolutionary history. As a result, we confirmed that the virus was a "swine-human-avian" reassortant virus.

2. Materials and Methods

2.1. Materials

Materials and Reagents. The lungs and lymph nodes used in the experiment were collected from a large-scale swine farm suspected to be infected with swine influenza in Fujian Province, China. EDS-76 and ND were positive sera provided by Fujian Institute of Animal Husbandry and Veterinary Medicine; the standard antigens and its antisera of H1, H3, H5, H7, and H9 subtypes and antisera of N1, N2, and N9 subtypes were purchased from Harbin Veterinary Research Institute; RNA extraction kit Trizol, plasmid extraction kit, LATaq DNA polymerase, RNasin Inhibition, M-MLV, dNTP, BamHI, HindIII, and PMD18-T vector were purchased from Dalian Bao Biological Company; *Escherichia coli* engineering strain DH5a was preserved by our laboratory.

Test Animals. BALB/c mice were collected from Fujian Medical Laboratory Animal Center; 5-week-old piglets were purchased from a large-scale pig farm in Fuzhou, which were confirmed to be negative in serological antibody examination; SPF eggs were purchased from Fuzhou Dabei Agricultural Biotechnology Co. and incubated to 9 to 11 days of age.

2.2. Virus Isolation

Acquisition and Processing of Disease Materials. The lung and lymph node tissues of pigs suspected to be infected with swine influenza were cut into pieces and ground into homogenate under sterile conditions. Hank's solution was added to it to make 10% suspension solution, which was then undergone frozen and thawed for three times, centrifuged at 12000 rpm for 10 min. The supernatant was filtered and sterilized, and placed at −70°C freezer for further use.

Virus Isolation and Culture. The supernatant was inoculated into allantoic cavity of 9 to 11-day-old SPF chicken embryos with 0.1 mL/embryo, and each sample was inoculated into 4 embryos. The inoculated embryos were incubated in an incubator at 37°C and a relative humidity of 55%. The eggs were candled twice a day to inspect the vitality of chicken embryos for 6 days. The embryos died within 24 hours were abandoned, whereas those died between 24 and 96 hours were stored overnight at 4°C. Allantoic fluid was collected and the hemagglutination activity was detected. If the hemagglutination result is positive, we preserved chicken embryo allantoic fluid; otherwise the allantoic fluid were blindly passed for three generations and discarded if the result was still negative.

2.3. Virus Identification

Swine Influenza RT-PCR. A pair of primers was designed according to the M gene conserved region of swine flu registered in GenBank, i.e., SIV-F: 5′-CAAGACCAATCCTGTCACCTC-3′, and SIV-R: 5′-AAGACGATCAAGAATCCAC AA-3′. It was then amplified into a fragment of 684 bp by synthetic company Dalian Biological Engineering Technology. The viral RNA was extracted by conventional methods. The 10.0 μL RT reaction system consists of DEPC water (1.25 μL), 5 × AMV Buffer (2.0 μL), 25 mg $MgCl_2$ (1.0 μL), dNTP mixture (1.0 μL), RNase inhibitor (0.25 μL) of AMV (0.5 μL), random primer (1.0 μL), and RNA sample template (3.0 μl). The RT reaction conditions were 30°C for 10 min, 42°C for 1 h, and 99°C for 5 min. The 25.0 μL PCR reaction system consists of Nuclease-Free Water (10.0 μL), GoTaq Green Master Mix (10.0 μL), upstream primer (0.5 μL), downstream primer (0.5 μL), and RT reaction product (4.0 μL). The PCR reaction conditions were 95°C for 5 min, 94°C for 1 min, 54°C for 1 min, 72°C for 1 min with 35 cycles, and then 72°C for 10 min and 4°C for preservation. After the reaction, the PCR products were electrophoresed on 1% agarose gel, and the positive product was sequenced by Bioengineering (Shanghai) Co., Ltd.

Electron Microscopy. After allantoic fluid was centrifuged by differential and density gradient method, the morphology and size of the virus were observed by transmission electron microscopy.

Agar Diffusion Test. The isolated chicken embryo allantoic fluid was purified, which can be used as antigen after adding SDS. After that, it was under agar diffusion test with standard serum. The results were collected at 24 h, 48 h, and 72 h, respectively.

2.4. Identification of Serum Subtypes. The HA subtype of swine influenza virus was identified by comparing the isolated virus with the standard positive sera of swine influenza by hemagglutination assay (HA) and hemagglutination inhibition assay (HI). Neuraminidase inhibition assay (NI) was used to identify NA subtypes.

Hemagglutination Assay (HA). We added 50 μl of virus antigen in the first well of each row and used a pipette to mix well; we then added 50 μl of the mixed solution to the second well, diluted consecutively until the 11th well. The 12th hole without antigen was taken as a control. After that, we added 50 μl 1% pigeon erythrocyte suspension to each well, oscillated using microoscillator for 1 min and recorded the results after 20 ~ 30 min quiescence in room temperature.

Hemagglutination Inhibition Assay (HI). According to the hemagglutination test, we made 8-unit and 4-unit antigen by adding physiological saline. We then added 50 μl 8-unit antigen into the first well and 50 μl 4-unit antigen into the 2nd, 3rd, . . ., and 11th well and 50 μl normal saline to the 12th well as negative control. After that, we took 50 μl standard positive serum to the first well, mixed them even, and took 50 μl to the second well. The process was repeated until the 11th well. After 15 min at room temperature, we added 50 μl 1% pigeon erythrocyte suspension, oscillated using microoscillator for 1 min and recorded the results after 20 min quiescence in room temperature. The hemagglutination inhibition titer of the serum was the maximum dilution of sera with 50% inhibition of erythrocyte agglutination. Hemagglutination inhibition titer ≥ 4log3 serum was determined as positive.

Neuraminidase Inhibition Assay (NI). The allantoic fluid of the virus was tested against the standard antiserum of three NA subtypes.

2.5. Half of the Chicken Embryo Infection Determination. The virus liquid of each dilution was inoculated into 9-day-old SPF chicken embryos with 4 pieces/group and 0.1 mL/piece. The embryos were then sealed with wax and cultured at 37°C for 7 days. The embryos which died within 24 hours were abandoned and allantoic fluid from embryos which died after 24 hours were collected for hemagglutination assay and the EID_{50} was calculated by Karber's method.

2.6. Identification of Viral Physical and Chemical Characteristics. Virus allantoic fluid was centrifuged at 3000 rpm for 15 min, and the supernatant was distributed into several tubes for further use. We then added SDS (1 mg/mL, pH 8.2 for 1 h), chloroform (final concentration 48 mL/L at 4°C for 10 min), ether (final concentration 200 mL/L at 4°C for 24 h), hydrochloric acid (final concentration 1 mL/L for 3 min), 1% potassium permanganate (for 5 min), formaldehyde (final concentration 0.2%), and 75% ethanol (for 5 min) and treated them with water at 56°C for 0,5, 10,15, 20, 30, 60, 90, 120, and 180 mins overnight. The treated virus was inoculated into 9-day-old SPF chicken embryo with 0.1 mL/embryo. Allantoic fluid or the embryo died in 24 ~ 96 h were collected for HA assay.

2.7. Animal Vaccination Test

BALB/c Mice Infection Assay. The virus with the dose $2 \times 10^5 EID_{50}$ was inoculated intranasally into ten 6-week-old BALB/c mice and physiological saline was inoculated into 10 mice as controls. One BALB/c mouse was dissected on the 1st, 3rd, 5th, 7th, 9th, 11th, and 13th days after inoculation. The lungs were harvested for virus isolation. The mental and diet status of the mice were also observed for 2 weeks.

Infection Assay on Piglets. Twelve 5-week-old piglets, which were tested negative for swine influenza virus, were divided into experiment and control group evenly. Piglets in experiment were inoculated intranasally with $2 \times 10^6 EID_{50}$ dose viruses and those in control group were inoculated with normal saline. The body temperature of all piglets was measured every day after inoculation, and the diet and mental status of piglets were observed continuously for 28 days.

2.8. Whole Gene Sequencing

Primer Synthesis. According to the H1N2 subtype of swine influenza virus in GenBank, 12 pairs of primers containing 8 gene fragments were designed by oligo 6.0 software. The primers were available upon request.

Virus RNA Extraction and RT-PCR Reaction. Total RNA was extracted from the viral RNA extraction kit Trizol. The RT reaction conditions were DEPC water (1.25 μL), MgCl2 (1.0 μL), 5 × MMLV Buffer (2.0 μL), dNTP Mixture (1.2 μL), MMLV (0.5 μL), RNase Inhibition (0.25 μL), Uni-12 primer (1.0 μL), RNA (3.0 μL); the reaction conditions were 30°C for 10 min, 42°C for 60 min, and 99°C for 5 min and preserved at 4°C. The PCR reaction conditions were 10× PCR Buffer (2.0 μL), dNTP Mixture (1.2 μL), DNA polymerase (0.2 μL), Up Primer (0.4 μL), Down Primer (0.4 μL), and template (2.0 μL) and filled to 20 μL by adding H_2O. The reaction conditions were as follows: predenaturation at 95°C for 5 min, denaturation at 94°C for 1 min, annealing at 53.5-61.0°C for 1 min, extension at 72°C for 1 min, and extension at 72°C for 10 min. The PCR products were identified by 1.0% agarose gel electrophoresis

Gene Cloning and PCR Identification. The PCR product was ligated into PMD18-T vector and transferred into DH5α strain. The recombinant plasmid was identified by 1.0% agarose gel electrophoresis, and the positive plasmid was sent to Bioengineering (Shanghai) Co., Ltd., for sequencing.

2.9. Comparison of Gene Sequence Analysis. The sequences were put into GenBank for homology search using BLAST, and the sequences of subtypes human influenza H3N2, swine flu H3N2, classical swine flu H1N1, and representative swine flu H1N2 were downloaded and analyzed by DNAStar software. The sequence alignments were done by MUSCLE software in MEGA6 (http://www.megasoftware.net/). The phylogenetic trees were generated using the Neighbor-Joining method with 1000 bootstrap replicates.

3. Results

3.1. Virus Isolation. We collected allantoic fluid from inoculated SPF chick embryos died between 24 to 72 hours, whose hemagglutination titer was determined to be 2^5. By autopsy, we found that there are serious embryo hemorrhage lesions.

3.2. Virus Identification. RT-PCR amplification products were detected by agarose gel electrophoresis (Figure 1). As can be seen, the amplified fragments are of sizes about 684 bp as expected. Sequence analysis showed that the gene fragment was derived from swine flu. By visualizing allantoic fluid after centrifugation using electron microscope, one can find influenza virus particles of diameter 80 ~ 120 nm. They are

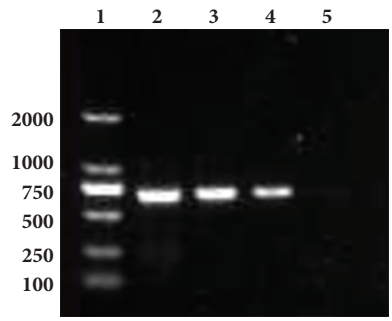


Figure 1: RT-PCR results of M gene. 1: marker 2000; 2, 3, and 4: M; 5: negative control.

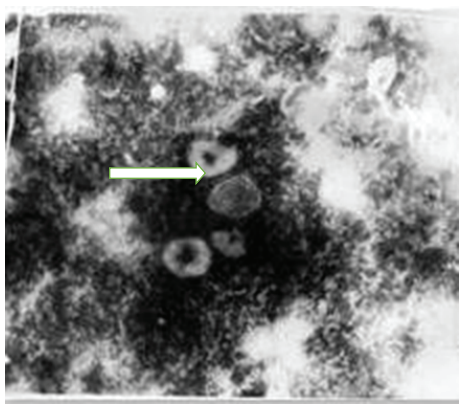


Figure 2: Electron specular photograph of allantoic fluid after centrifugation.

in spherical, horseshoe-shaped, and filamentous enveloped shapes (see Figure 2).

There are white precipitate line identified in the agar diffusion assay between tested antigen and standard influenza A serum, indicating that the isolated virus was an influenza A virus. We then performed HI assay between allantoic fluid and standard positive sera of EDS-76 and ND. The results showed that the isolated strain cannot be inhibited by both EDS-76 and ND, indicating that it is not an EDS or NDV virus. We also performed HA and HI assay against the standard serum of H1 subtype and showed that it can be inhibited by H1 serum with HI value 8 log2, suggesting that the hemagglutinin of the virus is of H1 subtype. In addition, the NI assay suggested that the neuraminidase of the virus is of subtype N2. As a result, we named the virus A/Swine/Fujian/F1/2010 (H1N2). Finally, the EID_{50} value of the fourth-generation allantoic fluid isolated from culture was $10^{-4.6}$/0.1 mL. The virus lacks resistance to SDS, chloroform, ether, hydrochloric acid, 1% potassium permanganate, formaldehyde, and 75% ethanol and can easily be inactivated. The titer of hemagglutination decreased by 2 titers at 56°C for 5 min. According to the method of determining the thermal stability, it can be determined that the virus strain is heat-labile.

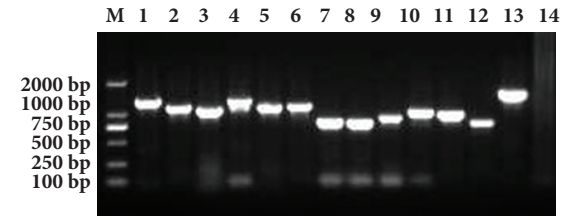


Figure 3: RT-PCR electrophoresis of A/Swine/Fujian/F1/10 (H1N2) strain. M: marker 2000; 1 and 2: PB2; 3 and 4: PB1; 5 and 6: PA; 7 and 8: NP; 9: NS; 10: M; 11 and 12: HA; 13: NA; 14: negative control.

3.3. Animal Vaccination. We first performed BALB/c mouse infection assay. The mice in the experiment group showed symptoms of mental exhaustion, coarse coat, and dyspnea symptoms 3 days after inoculation, but no mice died. The control group was normal. Influenza virus was isolated from the lungs of mice from 1 to 5 days after inoculation, and no virus could be isolated after 7 days after inoculation.

We then infected virus with 6-weed-old piglets. Similar to mice, piglets in experiment group developed symptoms of high body temperature, loss of appetite, cough, and runny nose after 4 days of inoculation, and the symptoms were relieved after 7 days until all the piglets were recovery. No piglets died in the process. In contrast, piglets in control group were normal.

3.4. Whole Genome Sequencing and Genetic Variation Analysis. The PB2, PB1, PA, HA, NP, NA, M, and NS genes of A/Swine/Fujian/F1/2010 (H1N2) strain were amplified by RT-PCR, respectively. After sequencing, we obtained 2341 bp, 2302 bp, 2190 bp, 1709 bp, 1543 bp, 1443 bp, 991 bp, and 882 bp gene fragments (Figure 3), which were uploaded to NCBI with accession numbers KM186312, KM186311, KM186310, KM186305, KM186308, KM186307, KM186306, and KM186309, respectively.

We then performed nucleotide homology and genetic phylogenetic analysis using BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi), DNAStar (https://www.dnastar.com/), and MEGA6 using Neighbor-Joining method with 1000 bootstrap replicates. We used K80 model to estimate the distances between nucleotides. Specifically, the eight gene fragments of A/Swine/Fujian/F1/2010 (H1N2) strain were compared with human influenza, avian influenza, and swine influenza virus deposited in NCBI flu database.

The nucleotide sequences of HA, NA, PB1, PA, PB2, NP, M, and NS genes all had the highest homology with A/Swine/Shanghai/1/2007 (H1N2) with identity 98.9%, 98.9%,99.0%, 98.6%, 99.0%, 98.9%, 99.3%, and 99.3%, respectively. According to the phylogenetic tree, the HA, NA, PB1, PA, PB2, NP, and M gene fragments of A/Swine/Fujian/F1/2010 (H1N2) is genetically close to those of A/Swine/Shanghai/1/2007(H1N2), while its NS fragment is close to A/Swine/Guangxi/13/2006 (H1N2). We illustrated the phylogenetic trees of HA, NA, PB1, PA, and PB2 in Figures 4–8, respectively.

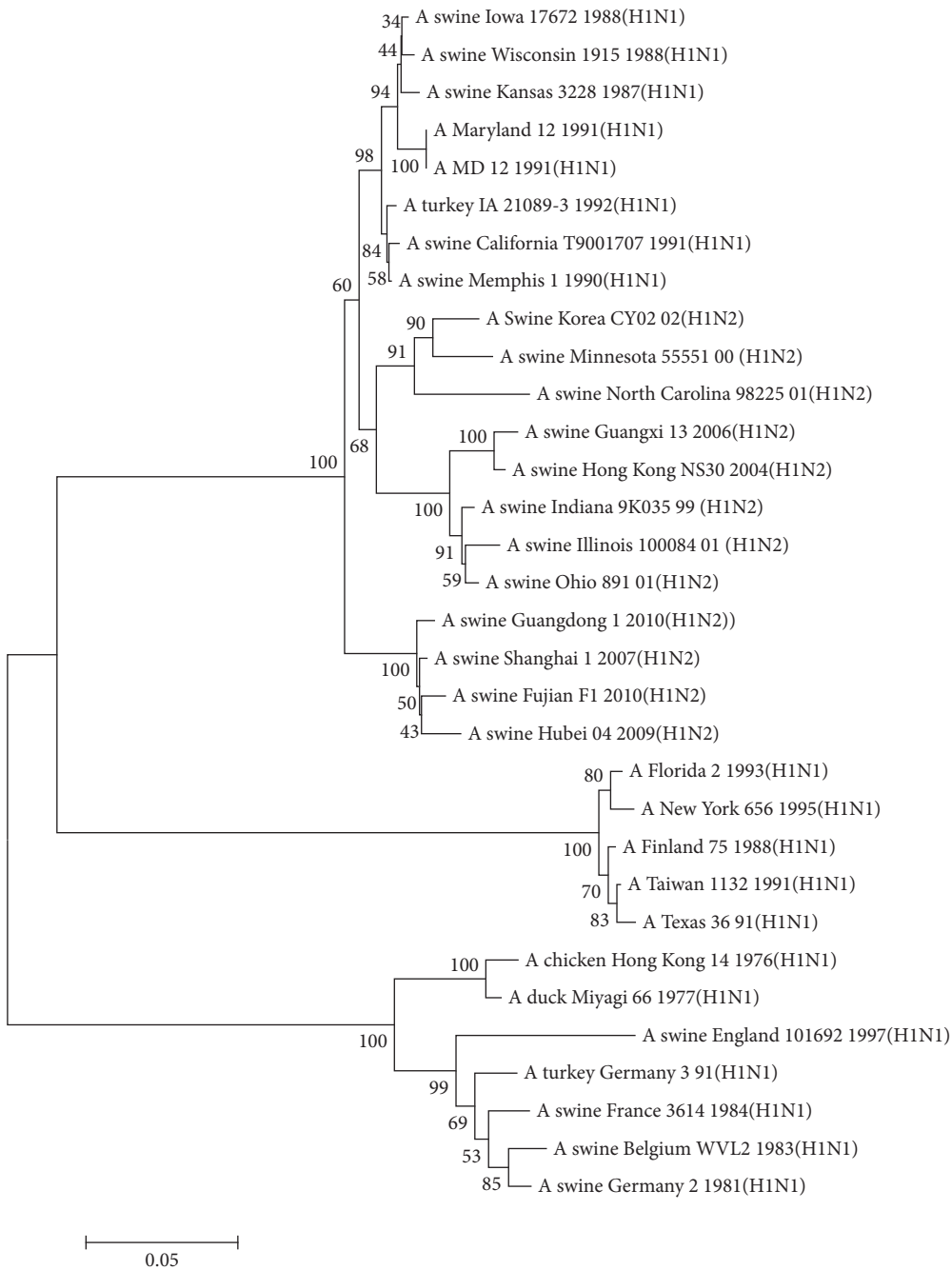


Figure 4: Phylogenetic tree based on HA nucleotide sequence.

3.5. Analysis on the Sources of Gene Fragments. The analysis of 8 gene fragments of A/Swine/Fujian/F1/2010 (H1N2) strain showed that the HA, NP, M, and NS gene fragments were originated from classical pig influenza H3N2 subtype, NA and PB1 were derived from human swine influenza H3N2 subtype, and PB2 and PA gene were from avian influenza virus, as shown in Figure 9. The results indicated that the strain was a "swine-human-avian" reassortant virus.

4. Discussion

Swine as a host of influenza virus can simultaneously infect different types and subtypes of influenza viruses. Therefore, most scholars believe that swine plays an important role in influenza pandemic as a "mixer" of virus reassortment [2]. When pigs infect both human and avian influenza viruses at the same time, new strains may be produced due to gene reassortment, causing epidemics. The southern part

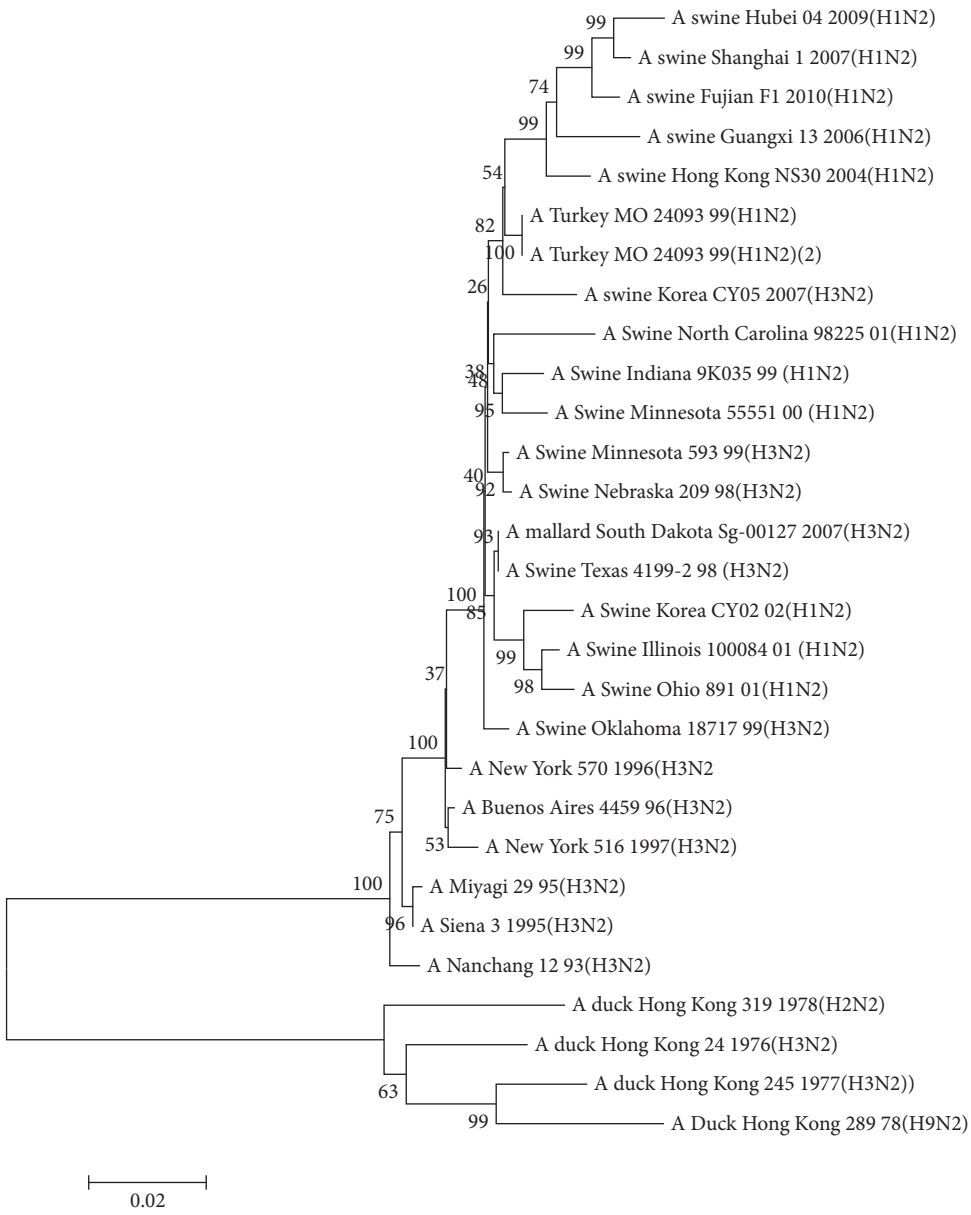


Figure 5: Phylogenetic tree based on NA nucleotide sequence.

of China is considered to be the source of the world's flu outbreak due to its unique ecological, geographical, and climatic conditions. Fujian locates in southern China, next to Guangdong, one of the most frequent influenza out-breaking places in the world. There are a lot of pig farms and waterfowl habitats in Fujian, greatly increasing the chance of influenza virus reassortment.

Currently, there are a lot of reports about swine influenza virus reassortment all over the world. For example, the H3N2 reassortant virus is isolated from North Carolina in 1998, whose HA, NA, and PB1 fragments are derived from human influenza H3N2 subtype virus; NP, NS, M, PB2, and PA were derived from classical swine influenza H1N1 viruses; and PB2 and PA were originated from H9N2 avian influenza viruses. In 2007, Yao et al. reported a two-source H3N2 reassortant swine influenza virus [16]. The M and NS fragments of the strain were derived from the H1N1 swine influenza virus, whereas HA and NA were derived from H3N2 human influenza virus. In 1994, Brown et al. obtained a swine-human-avian three-source recombinant H1N2 swine influenza virus [17]. The HA gene of the strain was derived from human influenza H1N1. The NA gene was derived from the porcine H3N2 subtype, while the other six genes were originated from the H1N1 flu virus of the avian species. Karasin et al. ([6]; Karasin et al. 2002) also isolated the "swine-human-avian" triple-reassortant H1N2

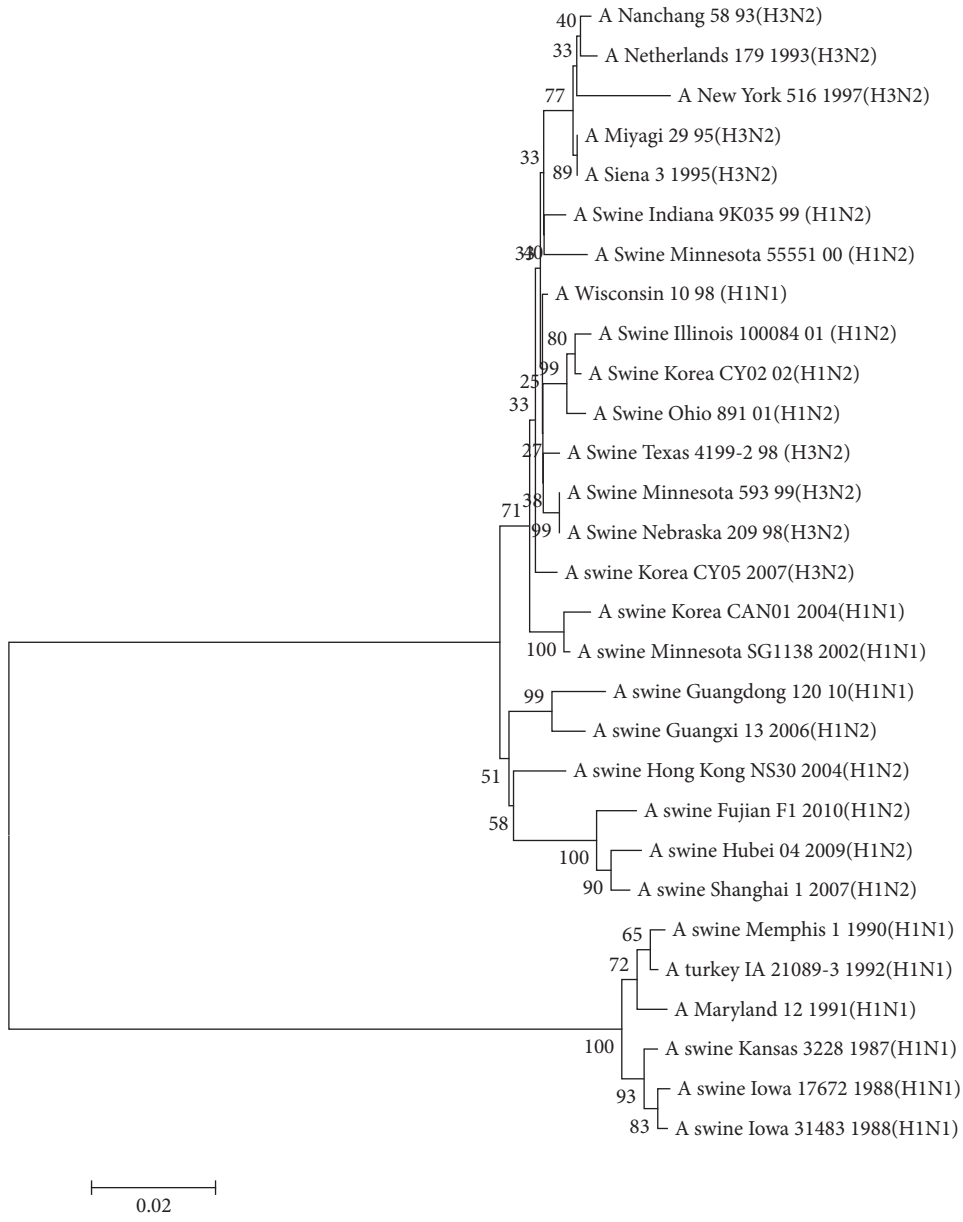


Figure 6: Phylogenetic tree based on PB1 nucleotide sequence.

swine influenza virus in the United States from 1999 to 2001. The HA, M, NP, and NS were originated from the classical H1N1 swine flu viruses, the NA and PB1 genes were from the human H3N2 influenza viruses, and the PB2 and PA genes are from the avian flu viruses.

In this study, a virus was isolated from the lymph nodes and lungs of pigs suspected to be infected with swine flu viruses. Our results showed that the virus had strong pathogenicity to chick embryo and could kill the chicken embryo in 48 hours. The allantoic fluid had hemagglutination activity. It is confirmed to be a H1N2 subtype virus by HA, HI, and NA assays, electron microscopy, and whole genome sequencing and thus named A/swine/Fujian/F1/2010 (H1N2). Based on the role of pigs as intermediate hosts in the SIV interspecies transmission chain and the complexity of influenza ecology in China, it is of great importance to understand the prevalence of influenza viruses circulating in swine herds. Although EID_{50} of the isolated strains was $10^{-4.6}$/0.1 mL, it could infect mice and piglets without killing them, of which the isolated strains were less pathogenic.

Based on the whole genome sequencing analyses, A/Swine/Fujian/F1/2010 (H1N2) was a "swine-human-avian" reassortant virus, highly similar to A/Swine/shanghai/1/2007 (H1N2), another reassortant virus isolated in Shanghai, China. We inferred that the HA, NP, M, and NS genes were

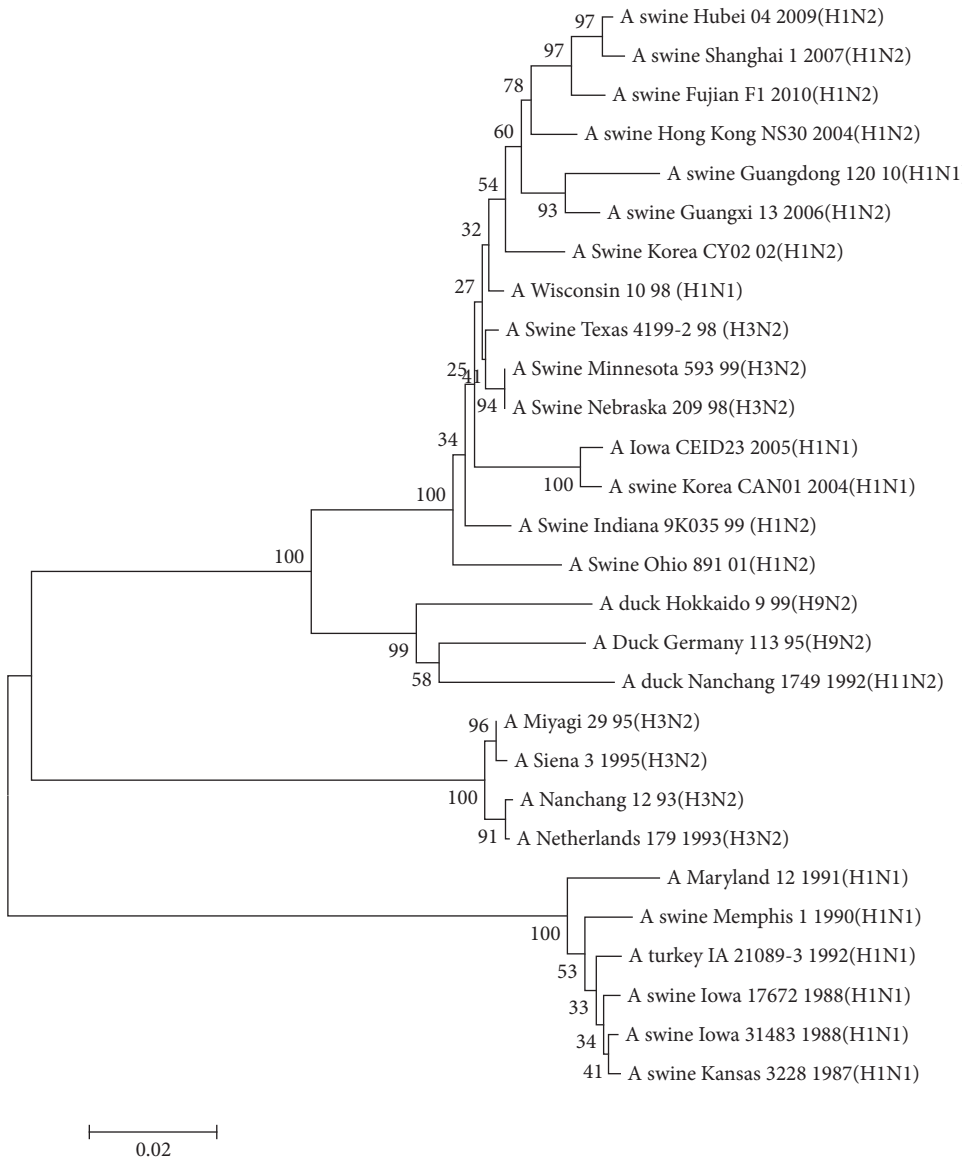


Figure 7: Phylogenetic tree based on PA nucleotide sequence.

derived from the swine classical H1N1 influenza virus. The NA and PB1 genes were derived from human H3N2 subtype influenza viruses and PA and PB2 genes were from the avian influenza virus. The genetic characteristics of swine influenza virus isolation show that swine does play an important role as a "mixer" during the spread of influenza virus. Studies on the molecular and genetic evolution of swine flu are of great importance for timely detection of swine flu. The surveillance of the variation of swine influenza is critical for prevention and control of flu outbreaks.

Conflicts of Interest

The authors have declared no conflicts of interest.

Authors' Contributions

Lun-Jiang Zhou conceived and designed the experiments. Long-Bai Wang performed the experiments and analyzed the data. Long-Bai Wang and Lun-Jiang Zhou wrote the paper. Cheng-Yan Wang, Xue-Min Wu, Yong-Liang Che, Ru-Jing Chen, and Lun-Jiang Zhou contributed to the discussion and helped to revise the paper. All authors have reviewed and approved the final manuscript.

Acknowledgments

This work was partially supported by the Public Welfare Scientific Research Projects of Fujian Province (no. 2009K10025-5) and the Modern Agricultural Pig Industry Technology System Project of Fujian Province (2014-2017).

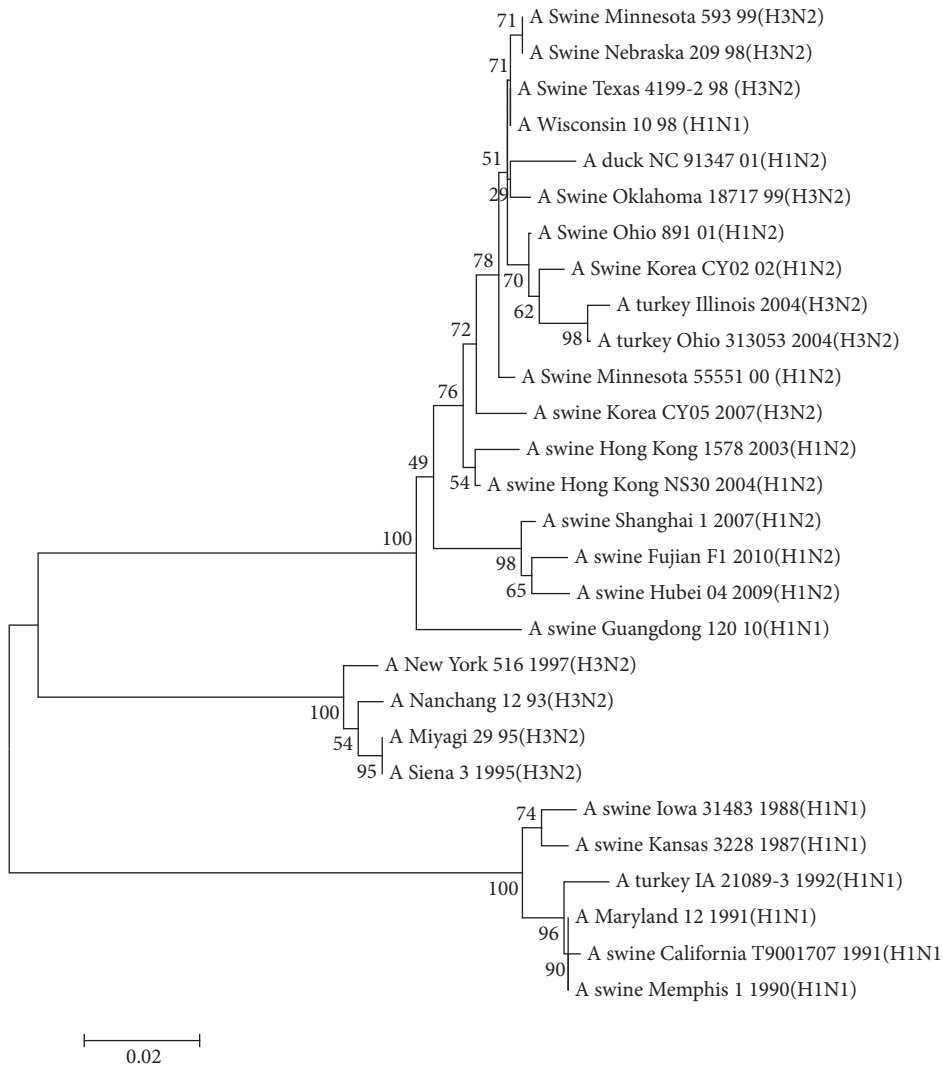


Figure 8: Phylogenetic tree based on PB2 nucleotide sequence.

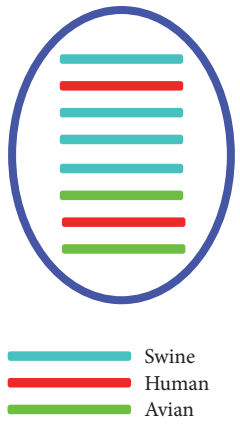


Figure 9: Schematic diagram representing genes of A/Swine/Fujian/F1/2010(H1N2). The eight gene fragments were HA, NA, M, NP, NS, PA, PB1, and PB2 from top to bottom.

References

[1] H. Kothalawala, M. J. M. Toussaint, and E. Gruys, "An overview of swine influenza," *Veterinary Quarterly*, vol. 28, no. 2, pp. 45–53, 2006.

[2] R. G. Webster, W. J. Bean, O. T. Gorman, T. M. Chambers, and Y. Kawaoka, "Evolution and ecology of influenza A viruses," *Microbiology and Molecular Biology Reviews*, vol. 56, no. 1, pp. 152–179, 1992.

[3] N. N. Zhou, D. A. Senne, J. S. Landgraf et al., "Genetic reassortment of avian, swine, and human influenza A viruses in American pigs," *Journal of Virology*, vol. 73, no. 10, pp. 8851–8856, 1999.

[4] F. H. Top and P. K. Russell, "Swine influenza a at fort dix, new jersey (January–february 1976). iv. summary and speculation," *The Journal of Infectious Diseases*, vol. 136, pp. S376–S380, 1977.

[5] J. S. M. Peiris, Y. Guan, D. Markwell, P. Ghose, R. G. Webster, and K. F. Shortridge, "Cocirculation of avian H9N2 and

contemporary "human" H3N2 influenza A viruses in pigs in southeastern China: Potential for genetic reassortment?" *Journal of Virology*, vol. 75, no. 20, pp. 9679–9686, 2001.

[6] A. I. Karasin, M. M. Schutten, L. A. Cooper et al., "Genetic characterization of H3N2 influenza viruses isolated from pigs in North America, 1977-1999: Evidence for wholly human and reassortant virus genotypes," *Virus Research*, vol. 68, no. 1, pp. 71–85, 2000.

[7] L.-Y. Chang, S.-R. Shih, P.-L. Shao, D. T.-N. Huang, and L.-M. Huang, "Novel Swine-origin Influenza Virus A (H1N1): The First Pandemic of the 21st Century," *Journal of the Formosan Medical Association*, vol. 108, no. 7, pp. 526–532, 2009.

[8] C. Scholtissek, W. Rohde, V. Von Hoyningen, and R. Rott, "On the origin of the human influenza virus subtypes H2N2 and H3N2," *Virology*, vol. 87, no. 1, pp. 13–20, 1978.

[9] Y. Kawaoka, S. Krauss, and R. G. Webster, "Avian-to-human transmission of the PB1 gene of influenza A viruses in the 1957 and 1968 pandemics," *Journal of Virology*, vol. 63, no. 11, pp. 4603–4608, 1989.

[10] I. H. Brown, "The epidemiology and evolution of influenza viruses in pigs," *Veterinary Microbiology*, vol. 74, no. 1-2, pp. 29–46, 2000.

[11] R. J. Webby, S. L. Swenson, S. L. Krauss, P. J. Gerrish, S. M. Goyal, and R. G. Webster, "Evolution of swine H3N2 influenza viruses in the United States," *Journal of Virology*, vol. 74, no. 18, pp. 8243–8251, 2000.

[12] Y. L. Cong, J. Pu, Q. F. Liu et al., "Antigenic and genetic characterization of H9N2 swine influenza viruses in China," *Journal of General Virology*, vol. 88, no. 7, pp. 2035–2041, 2007.

[13] Q. Zhu, H. Yang, W. Chen et al., "A naturally occurring deletion in its NS gene contributes to the attenuation of an H5N1 swine influenza virus in chickens," *Journal of Virology*, vol. 82, no. 1, pp. 220–228, 2008.

[14] A. Moreno, I. Barbieri, E. Sozzi et al., "Novel swine influenza virus subtype H3N1 in Italy," *Veterinary Microbiology*, vol. 138, no. 3-4, pp. 361–367, 2009.

[15] W. D. Kundin, "Hong Kong A-2 influenza virus infection among swine during a human epidemic in taiwan," *Nature*, vol. 228, no. 5274, p. 857, 1970.

[16] Y. Yao, G.-H. Zhang, W.-J. Liu, T.-Q. Chen, and L. Sun, "Genome sequence analysis of an H3N2 subtype swine influenza virus isolated from Guangdong province in China," *Wei sheng wu xue bao = Acta microbiologica Sinica*, vol. 47, no. 5, pp. 805–809, 2007.

[17] I. H. Brown, P. Chakraverty, P. A. Harris, and D. J. Alexander, "Disease outbreaks in pigs in Great Britain due to an influenza A virus of H1N2 subtype," *Veterinary Record*, vol. 136, no. 13, pp. 328-329, 1995.

Hindawi
BioMed Research International
Volume 2018, Article ID 9321270, 7 pages
https://doi.org/10.1155/2018/9321270

Research Article

Abnormal Liver Function Induced by Space-Occupying Lesions Is Associated with Unfavorable Oncologic Outcome in Patients with Colorectal Cancer Liver Metastases

Zheng Jiang,[1] Chunxiang Li,[2] Zhixun Zhao,[3] Zheng Liu,[1] Xu Guan,[1] Ming Yang,[1] Xiaofu Li,[4] Dawei Yuan,[5] Songbo Qiu,[6] and Xishan Wang[1]

[1]*Department of Colorectal Surgery, Cancer Institute & Hospital, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing 100021, China*
[2]*Department of Thoracic Surgery, Cancer Institute & Hospital, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing 100021, China*
[3]*Department of Colorectal Cancer Surgery, The 2nd Affiliated Hospital, Harbin Medical University, Harbin, Heilongjiang 150001, China*
[4]*Department of Magnetic Resonance Imaging, The 2nd Affiliated Hospital, Harbin Medical University, Harbin, Heilongjiang 150001, China*
[5]*Genesis (Beijing) Co., Ltd., Beijing 100102, China*
[6]*Department of Experimental Therapeutics, The University of Texas MD Anderson Cancer Center, Houston, TX 77054, USA*

Correspondence should be addressed to Xishan Wang; wxshan1208@126.com

Received 29 November 2017; Revised 25 January 2018; Accepted 27 February 2018; Published 15 May 2018

Academic Editor: Tao Huang

An early prediction of prognosis for patients with colorectal liver metastasis (CRLM) may help us determine treatment strategies. Liver function reflects the effect of the overall metastatic burden. We investigated the prognostic value of liver function in CRLM patients. In our study, patients with abnormal LFTs (liver function tests) had a poorer prognosis than did those with normal LFTs ($P < 0.05$). A multivariate analysis revealed that LFTs was an independent prognostic factor for CRLM. For those patients with abnormal LFTs, novel prognostic contour maps were generated using LFTs, and no positive correlation exists between the values of survival duration and abnormal LFTs. Additionally, the MTVR (metastatic tumor volume ratio) was measured directly by magnetic resonance imaging and was shown to be highly correlated to LFTs by a Pearson correlation analysis. A multivariate logistic regression analysis also demonstrated that the MTVR and hepatectomy were independently predictive of abnormal LFTs. The space-occupying effect of metastatic lesions can cause abnormal LFTs, resulting in a poor prognosis. Biochemical analyses of LFTs at the initial diagnosis of CRLM enable the stratification of patients into low- and high-risk groups; it may help clinicians determine promising treatment strategies.

1. Introduction

More than 140,000 patients are diagnosed with colorectal cancer (CRC) each year in the United States [1]. Approximately 60% will develop liver metastases [2]. The prognosis for metastatic CRC has significantly improved in the past 10–15 years, with more effective surgical approaches and efficacious chemotherapy regimens making it possible for patients to undergo surgical resection [3]. Even though the vast majority of metastatic CRC patients (80%–90%) present with unresectable disease, modern combination chemotherapy results in a median survival duration of roughly 20 months [4–6]. However, evaluating the prognosis of patients with CRC liver metastasis (CRLM) is still challenging, and the results will influence treatment strategies.

The TNM Classification of Malignant Tumors is the main prognostic tool used in clinical practice. However, it is not sufficient to differentiate the likelihood of survival in stage IV

cases. Therefore, new methods of predicting and improving outcome are being explored [7–10]. Numerous oncologists and clinical researchers have assessed the relevance of liver function tests (LFTs) for early detection of liver metastasis in patients with different types of cancer [11, 12]. However, the conclusions remain inconsistent [13, 14].

Although the role of LFTs in identifying metastases to the liver remains unclear, on the basis of our clinical experience, we speculated that abnormal LFTs might be useful for predicting prognosis in patients with CRLM, which is induced by space-occupying lesions. Therefore, in this retrospective study, we determined the prognostic value of LFTs in patients with a definitive diagnosis of CRLM.

2. Materials and Methods

2.1. Patients. After receiving approval from the institutional review board, we searched the patient data bank at our institution to identify all consecutive patients who underwent operative and conservative treatment for CRLM and were first seen between December 1987 and June 2010. For compatibility, only patients who met the following criteria were considered for further analysis: (1) age ≥ 18 years and ≤75 years; (2) none of the patients in either group had other known liver disease at entry into the study; (3) previous normal LFTs; (4) no evidence of extrahepatic metastases; (5) no history of cancer; (6) complete follow-up data; and (7) LFTs available for the date of diagnosis of CRLM. Most patients with resectable liver metastases from colorectal cancer have received operation within 1 week after initial diagnosis.

The diagnosis of liver metastasis was confirmed by fine needle aspiration biopsy or typical clinical and imaging findings, disease progression, and the absence of any additional cancer. Tumors were staged in accordance with the American Joint Committee on Cancer Staging System. For the survival analysis, progression-free survival was measured from the time of diagnosis to the time of tumor progression. Overall survival (OS) was defined as the time from the date of diagnosis of liver metastases to the date of patient death or last follow-up.

2.2. Biochemical Measurements. LFTs were measured in a core laboratory and were considered abnormal when levels exceeded 40 U/l for alanine transaminase (ALT), 40 U/l for aspartate transaminase (AST), 60 U/l for gamma glutamyltransferase (γGT), 240 U/l for lactate dehydrogenase (LDH), and 150 U/l for phosphatase alkaline (AP). The blood was taken at inpatient for most patents with simultaneous hepatic metastases. 125 patients who developed liver metastases while undergoing regular (every 3 months) follow-up with LFTs and liver imaging were included in this study, and the blood was taken at outpatient clinic. The blood was taken at the initial diagnosis for CRLM, and the value was analyzed in this study.

The LFTs were classified as normal or abnormally elevated, according to the laboratory ranges. They were analyzed in isolation or were combined. Combined tests were analyzed using the following variables: 1 abnormal test result, 2 abnormal test results, 3 or 4 abnormal test results, or 5 abnormal test results.

2.3. Magnetic Resonance Imaging (MRI) Measurements. One hundred thirteen patients underwent multiphase liver MRI at the initial diagnosis of CRLM. MRI examinations were conducted using a 1.5 T system (Magnetom Symphony, Siemens, Erlangen, Germany). Volumetry on MRI was performed by one investigator (XFL) with 12 years' experience in abdominal MRI, supervised by an experienced hepatobiliary surgeon. After the imaging data had been transferred, the volume of the liver and lesions were measured using ImageJ, a software package for image analysis developed by the National Institutes of Health that can be freely downloaded from their website (https://rsb.info.nih.gov/ij/download.html).

All calculations using total liver volumes and metastatic tumor volumes (MTVs) were performed without liver remnant volumes. The MTV ratio (MTVR) was calculated as MTV/total liver volume.

2.4. Statistical Analyses. Patients' clinical characteristics are presented as means or medians for continuous variables and as percentages for categorical variables. Comparisons between normal and abnormal LFT groups for categorical variables were performed using Fisher's exact test or χ^2 tests. Survival curves were constructed using the Kaplan-Meier method, and the univariate survival difference was determined using the log-rank test. Time-point survival was estimated using the life-table method. Adjusted hazard ratios (HRs) with 95% confidence intervals (CIs) were calculated using Cox proportional hazards models.

The correlation between MTVR and LFTs was measured using the Pearson's R correlation test. A multivariate logistic regression analysis was performed to determine the variables associated with abnormal LFTs. Statistical analyses were performed using Stata statistical software, version 10.0 (StataCorp). Two-sided $P < 0.05$ was considered statistically significant.

3. Results

3.1. Characteristics of the Study Population. Of 1337 patients with CRLM, 552 (349 men and 203 women) met the inclusion criteria and were considered for further analysis; their median age was 58 years (range: 22–75). Their demographic and primary tumor characteristics are summarized in Table 1. The primary tumor was located in the colon in 271 patients (49.1%) and the rectum in 281 patients (50.9%). Primary cancer resection was performed in 450 patients (81.5%). Two hundred eighty-six patients (51.8%) had positive lymph nodes, as determined by a pathologic analysis of the colorectal specimen. Hepatic metastases were diagnosed simultaneously in 427 patients and metachronously in 125 patients and developed a mean of 20.9 months after CRC resection. Among the 552 patients with hepatic metastases from CRC, 22 (4.0%) underwent metastasectomy and 79 (14.3%) underwent ablation. Chemotherapy was administered in 193

Table 1: Demographic and primary tumor characteristics.

Clinicopathological features	Number (%)
Number of patients	552
Median age at diagnosis (range)	58 (22–87)
Gender	
Female	203 (36.8)
Male	349 (63.2)
Age	
≤60 years	313 (56.7)
>60 years	239 (43.3)
Location	
Rectum	281 (50.9)
Colon	271 (49.1)
Differentiation	
Well	13 (2.4)
Moderate	331 (60.0)
Poor	208 (37.7)
Mucinous histology	
Yes	36 (6.5)
No	516 (93.5)
T classification[&]	
T1/T2	24 (4.3)
T3/T4	368 (66.7)
Unknown	58 (12.9)
N classification[&]	
N0	160 (29.0)
N1	139 (25.2)
N2	147 (26.6)
Unknown	4 (0.7)
Perioperative chemotherapy	
Yes	397 (71.9)
No	155 (28.1)
Perioperative radiotherapy	
Yes	341 (61.8)
No	211 (38.2)
Resection margin[*]	
R0	427 (77.4)
R1	23 (4.2)

[*] Defined by findings on final pathological analysis (microscopic and major).
[&] American Joint Committee on Cancer Staging System.

patients (35.0%), with 9 patients treated by hepatic artery infusion. Ten patients (1.8%) underwent radiotherapy, and 384 received best supportive care (69.6%). The minimum follow-up duration was 0 months, and the maximum was 213 months (mean ± standard deviation, 13.5 ± 15.6 months).

3.2. Survival Differences by LFTs. Table 2 displays the detailed survival characteristics according to the LFTs. Patients with abnormal LFTs had poorer OS and progression-free survival durations than did those with normal LFTs. Among patients with abnormal values, those with elevated LDH levels had the poorest prognosis, with a median survival duration of

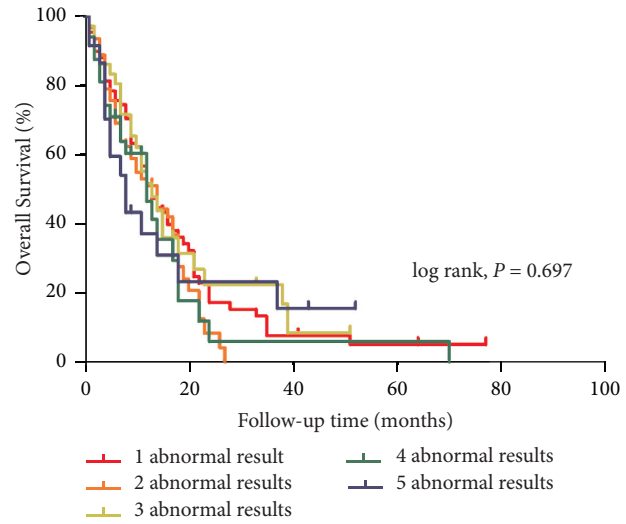


Figure 1: Kaplan-Meier curves showed no survival difference between the subgroups with isolated and combined variables.

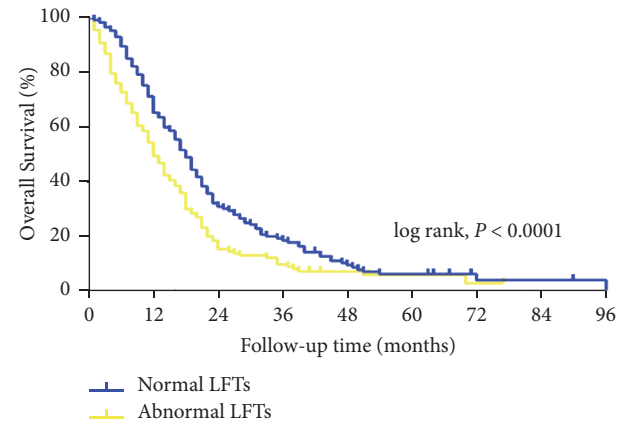


Figure 2: Kaplan-Meier curves showed significant survival difference between two subgroups in accordance with LFTs values.

only 9 months. The 2-year survival rates of the five subgroups (normal, AP, γGT, LDH, and ALT and/or AST) were 31.0%, 20.0%, 17.8%, 8.0%, and 16.8%, respectively ($P < 0.05$). When the combined LFTs were analyzed, however, there was no significant difference in prognosis between the patients with isolated and combined variables (Figure 1). Therefore, we divided patients into two subgroups—abnormal and normal LFTs—and performed a survival analysis. As shown in Figure 2, there were significant differences in OS between the two subgroups ($P < 0.05$). The median survival durations of the abnormal and normal LFT subgroups were 12 and 18 months, respectively. On multivariate analysis using the Cox proportional hazard model, LFTs were also an independent prognostic factor for CRLM ($P = 0.0001$; HR with 95% CI: 1.52 [1.22–1.88]). Subsequently, for revealing the relationships between the values of LFTs and survival time, the prognostic contour maps were produced using abnormal LFTs values, which indicated the probability of outcome (Figure 3). We

Table 2: Colorectal liver metastases, treatment, and survival characteristics depending on normal/abnormal values of liver function tests.

Clinicopathological features	Normal	Abnormal AP	*P* value*	Abnormal γGT	*P* value*	Abnormal LDH	*P* value*	Abnormal ALT and/or AST	*P* value*
No. patients	257	92		204		160		155	
Largest tumor size (cm)									
≤5	153	21		77		54		58	
>5	20	27	**<0.001**	53	**<0.001**	49	**<0.001**	40	**<0.001**
Number of liver metastases									
1	65	13		37		26		20	
>1	166	73	**0.017**	157	**0.029**	128	**0.011**	125	**0.001**
Hepatectomy									
Wedge resection	14	1		2		2		2	
Segmentectomy	2	0		1		0		1	
Hemihepatectomy	0	0		1		0		0	
No	241	91	**0.049**	200	**0.026**	158	**0.015**	152	**0.044**
Type of ablation									
RFA	26	11		34		16		26	
Cryotherapy	6	0		1		1		1	
No	225	81	0.901	169	0.155	143	0.573	88	**0.007**
Chemotherapy									
Yes	81	34		71		57		64	
No	176	58	0.341	133	0.456	103	0.386	91	**0.044**
MoAbs									
Yes	2	1		2		0		0	
No	255	91	0.863	202	0.816	160	0.263	155	0.271
HAIP placement									
Yes	4	2		4		4		3	
No	253	90	0.696	200	0.741	156	0.495	152	0.773
Radiotherapy									
Yes	7	0		2		1		1	
No	250	92	0.110	202	0.179	159	0.129	154	0.139
Progression-free survival (months)									
≤12	225	72		165		129		118	
>12	27	17	**0.042**	33	0.065	27	0.056	34	**0.002**
Overall survival									
median (months)	18	10		14		9		12	
Hazard ratio (95% CI)	-	0.57 (0.33–0.70)		0.68 (0.51–0.83)		0.50 (0.32–0.56)		0.62 (0.40–0.76)	
P value (log-rank)	-	**<0.001**		**<0.001**		**<0.001**		**<0.001**	
Survival rate (%) (95% CI)									
3 months	96.6 (93.3–98.3)	83.5 (72.8–90.3)		85.7 (79.5–90.0)		84.8 (77.6–89.8)		88.7 (81.9–93.0)	
6 months	89.6 (84.9–92.9)	64.2 (51.6–74.2)		71.7 (64.3–77.9)		65.8 (57.1–73.2)		78.3 (70.2–84.5)	
12 months	65.4 (58.5–71.5)	42.5 (30.1–54.3)		51.5 (43.3–59.0)		36.7 (28.1–45.2)		55.7 (46.2–64.3)	
24 months	31.0 (24.0–38.2)	20.0 (9.8–32.8)		17.8 (11.2–25.8)		8.0 (3.6–14.6)		16.8 (9.1–26.5)	

*Compare with normal group. RFA, radiofrequency ablation; HAIP, hepatic artery infusion pump; MoAbs, monoclonal antibodies; CI, confidence intervals.

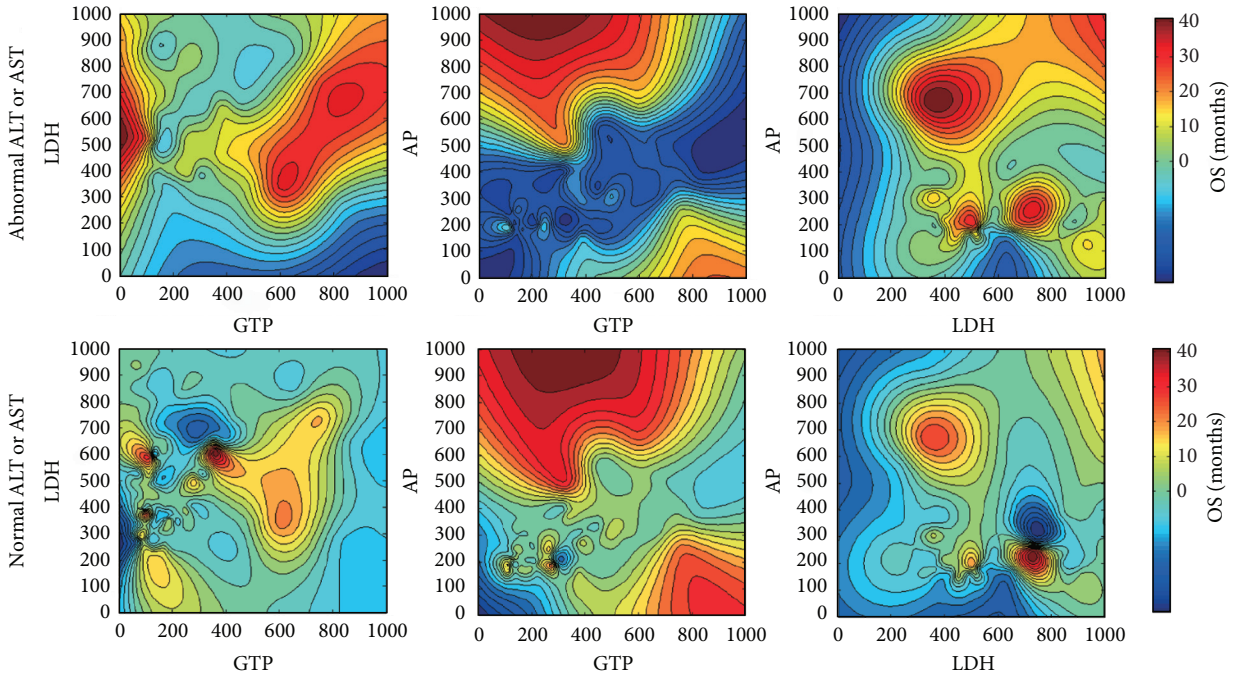


Figure 3: Contour maps for investigating the association between the values of survival duration and abnormal LFTs. Red areas depict favorable prognosis and blue areas unfavorable prognosis.

found that no positive correlation exists between the values of survival duration and abnormal LFTs.

3.3. A High MTVR Was Associated with Abnormal LFTs. Table 2 shows that LFTs were associated with marked clinical characteristics. Tumor size, number of liver metastases, progression-free survival, OS, hepatectomy or ablation history, and chemotherapy history were significantly associated with abnormal LFTs ($P < 0.05$). Larger number or size of liver metastatic tumors was the indicator of abnormal LFTs. We speculated that LFTs reflect the combined effect of diminished liver function and the overall metastatic burden.

The MTVR, which reflects the tumor burden, is initially expressed as a ratio of the metastatic tumor volume to the total liver volume, as measured directly by MRI. MTVR values were measured by retrospectively analyzing diagnostic MRI scans in the 113 CRLM patients for whom MRI data were available. A Pearson correlation analysis revealed that the MTVR was highly correlated with LFTs, such as phosphatase alkaline (Pearson correlation coefficient, 0.92) (Figure 4). We used a multivariate logistic regression analysis to estimate the parameters of a qualitative response model. Only the MTVR ($P < 0.001$, HR with 95% CI: 2.532 [1.410–4.545]) and hepatectomy ($P = 0.009$, HR with 95% CI: 3.448 [1.209–9.901]) had independent predictive value.

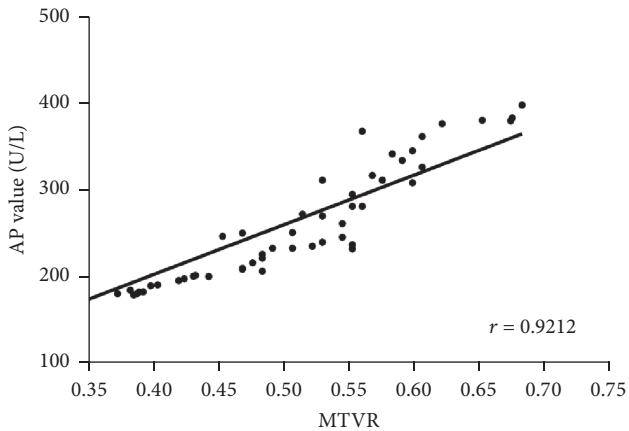


Figure 4: Correlation between MTVR and AP value.

4. Discussion

Worldwide, CRC is the second most commonly diagnosed cancer in women and the third in men, with over 1.2 million new cases and 608,700 deaths yearly [15, 16]. The liver is a common site of tumor spread, and in approximately 30% of cases, synchronous liver disease is present at the time of diagnosis. A further 50% of patients develop CRLM during the course of their illness [17]. Aggressive metastasectomy, combined with advancements in systemic chemotherapy, has led to significantly improved outcomes in patients with CRLM [18–20]. The overall 5-year survival rate for patients with resectable CRLM isolated to the liver is between 35% and 60% [21, 22] and the median survival duration is up to 22 months with systemic chemotherapy alone [23–25]. Unfortunately, the vast majority (75%–80%) of patients with CRLM are deemed unsuitable for surgical resection at initial diagnosis [26, 27]; thus, there remains a high demand for effective CRLM treatments and improved palliative results.

For longer life expectancy and better quality of life, evaluating the prognosis of CRLM patients is extremely important in clinical routine, and it may help us determine promising treatment strategies.

LFTs likely reflect the combined effect of diminished liver function and the overall metastatic burden. Certain types of agents used for chemotherapy, such as irinotecan, oxaliplatin, and 5-fluorouracil, can cause hepatic damage [28–30]. Most CRC patients in our study underwent adjuvant chemotherapy to decrease the risk of recurrence. To eliminate the effect of chemotherapy or other factors, only the patients with previously normal LFTs were included. We found that patients with abnormal LFTs had poorer outcomes than did those with normal LFTs on both univariate and multivariate survival analysis. We hypothesized that abnormal LFTs, which are caused by "space-occupying effect" of metastatic liver lesions, are associated with poor prognosis. For major hepatectomy, 30% of the total liver volume is considered to be sufficient to maintain adequate liver function [31]. In other words, up to 70% of the liver can be resected, including metastatic cancer and paracancerous and noncancerous tissue. Moreover, preexisting liver lesions may be underestimated on the basis of MRI volumetric studies. All of the above reasons may explain why some abnormal LFT values were associated with low MTVRs in this study (Figure 4). Although we found that the patients with abnormal LFTs had a poorer prognosis based on current data, no positive correlation between the values of survival duration and abnormal LFTs was observed in further analysis (Figure 3), which suggested caution in the exploration of treatment strategies for these patients.

MTVR serves as an alternative to two factors, tumor size and number of liver metastases in easier assessment of "space-occupying effect." Imaging plays an integral role in monitoring the status of CRLM. A variety of imaging techniques, including ultrasonography, computed tomography, and MRI, are used. MRI offers superior soft tissue resolution; therefore, it has several advantages over computed tomography in tissue characterization and the evaluation of background liver parenchyma. Small lesions can be detected and characterized more confidently with MRI [32, 33]. Therefore, in this study, we used MRI to measure MTVR. To verify the relevance of MTVR to abnormal LFTs, we used Pearson's R correlation test. We found that the value of LFTs increased with the MTVR value. Furthermore, on multivariate logistic regression analysis, MTVR and hepatectomy had independent predictive value.

Posthepatectomy liver failure remains an important cause of morbidity and mortality after major liver resection [34, 35]. Conventional biochemical LFTs, as evaluated by a routine blood test analysis, remain widely used and form an indispensable part of most definitions of liver failure [36, 37]. Undergoing liver resection results in postoperative changes in LFTs, which is consistent with our findings. Moreover, Grąt et al. recently found that LFTs on postoperative day 1 after major CLRM resection were substantially associated with outcome [38].

Our study has several limitations. First, most of the patients did not undergo hepatectomy for liver metastases. At the time many of these patients were diagnosed with CRLM, there were numerous barriers to obtaining surgical treatment, including a lack of experienced surgeons and factors such as a lack of health insurance, long travel distances, low health literacy, low education levels, and language barriers that affect patients' ability to navigate the medical system. Second, only some patients underwent MRI to determine the MTVR. MRI was used as a part of a routine diagnostic imaging procedure after 1998 at our hospital. In addition, imaging data of partial patients who underwent MRI were not available.

On the basis of our findings, we conclude that abnormal LFTs, which are induced not only by treatment factors but also by the space-occupying effects of metastatic lesions, will allow us to stratify CRLM patients into poor- and good-prognosis groups. It might help clinicians determine promising treatment strategies. However, a methodic and prospective study is needed to confirm these results, especially in high-risk patients selected by molecular analysis.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Zheng Jiang, Chunxiang Li, and Zhixun Zhao equally contributed to this work.

Acknowledgments

This study was supported by the National Key Research and Development Program of the Ministry of Science and Technology of China (2016YFC0905303), CAMS Innovation Fund for Medical Sciences (CIFMS) (2016-I2M-1-001), and Beijing Science and Technology Program (D17110002617004).

References

[1] K. D. Miller, R. L. Siegel, and C. C. Lin, "Cancer treatment and survivorship statistics, 2016," *CA: A Cancer Journal for Clinicians*, vol. 66, no. 4, pp. 271–289, 2016.

[2] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2016," *CA: A Cancer Journal for Clinicians*, vol. 66, no. 1, pp. 7–30, 2016.

[3] S. Kopetz, G. J. Chang, M. J. Overman et al., "Improved survival in metastatic colorectal cancer is associated with adoption of hepatic resection and improved chemotherapy," *Journal of Clinical Oncology*, vol. 27, no. 22, pp. 3677–3683, 2009.

[4] J. Scheele, R. Stangl, and A. Altendorf-Hofmann, "Hepatic metastases from colorectal carcinoma: Impact of surgical resection on the natural history," *British Journal of Surgery*, vol. 77, no. 11, pp. 1241–1246, 1990.

[5] H. K. Sanoff, D. J. Sargent, M. E. Campbell et al., "Five-year data and prognostic factor analysis of oxaliplatin and irinotecan combinations for advanced colorectal cancer: N9741," *Journal of Clinical Oncology*, vol. 26, no. 35, pp. 5721–5727, 2008.

[6] E. P. Araujo-Mino, Y. Z. Patt, C. Murray-Krezan et al., "Phase II trial using a combination of oxaliplatin, capecitabine, and celecoxib with concurrent radiation for newly diagnosed resectable rectal cancer," *The Oncologist*, vol. 23, no. 1, pp. 2–e5, 2018.

[7] F. Sellier, E. Bories, C. Sibertin-Blanc et al., "Clinical outcome after biliary drainage for metastatic colorectal cancer: survival

analysis and prognostic factors," *Digestive and Liver Disease*, vol. 50, no. 2, pp. 189–194, 2018.

[8] T. Murakami, H. Kikuchi, H. Ishimatsu et al., "Tenascin C in colorectal cancer stroma is a predictive marker for liver metastasis and is a potent target of miR-198 as identified by microRNA analysis," *British Journal of Cancer*, vol. 117, no. 9, pp. 1360–1370, 2017.

[9] K. Sasaki, G. A. Margonis, K. Maitani et al., "The prognostic impact of determining resection margin status for multiple colorectal metastases according to the margin of the largest lesion," *Annals of Surgical Oncology*, vol. 24, no. 9, pp. 2438–2446, 2017.

[10] S. Sabour, "Prognostic prediction by liver tissue proteomic profiling in patients with colorectal liver metastases; Rule of thumb," *Future Oncology*, vol. 13, no. 13, pp. 1133-1134, 2017.

[11] I. Kaiserman, R. Amer, and J. Pe'Er, "Liver function tests in metastatic uveal melanoma," *American Journal of Ophthalmology*, vol. 137, no. 2, pp. 236–243, 2004.

[12] M. S. Rocklin, A. J. Senagore, and T. M. Talbott, "Role of carcinoembryonic antigen and liver function tests in the detection of recurrent colorectal carcinoma," *Diseases of the Colon & Rectum*, vol. 34, no. 9, pp. 794–797, 1991.

[13] F. Mouriaux, C. Diorio, D. Bergeron, C. Berchi, and A. Rousseau, "Liver function testing is not helpful for early diagnosis of metastatic uveal melanoma," *Ophthalmology*, vol. 119, no. 8, pp. 1590–1595, 2012.

[14] A.-T. Le and C.-W. D. Tzeng, "Does finding early recurrence improve outcomes, and at what cost?: cost effectiveness of surveillance," *Journal of Surgical Oncology*, vol. 114, pp. 329–335, 2016.

[15] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, and J. Lortet-Tieulent, "Global cancer statistics, 2012," *CA: A Cancer Journal for Clinicians*, vol. 65, no. 2, pp. 87–108, 2015.

[16] R. L. Siegel, K. D. Miller, S. A. Fedewa et al., "Colorectal cancer statistics," *CA: A Cancer Journal for Clinicians*, vol. 67, no. 3, pp. 177–193, 2017.

[17] E. Van Cutsem, B. Nordlinger, R. Adam et al., "Towards a pan-European consensus on the treatment of patients with colorectal liver metastases," *European Journal of Cancer*, vol. 42, no. 14, pp. 2212–2221, 2006.

[18] Y. Fong, A. M. Cohen, J. G. Fortner et al., "Liver resection for colorectal metastases," *Journal of Clinical Oncology*, vol. 15, no. 3, pp. 938–946, 1997.

[19] T. M. Pawlik, R. D. Schulick, and M. A. Choti, "Expanding criteria for resectability of colorectal liver metastases," *The Oncologist*, vol. 13, no. 1, pp. 51–64, 2008.

[20] T. L. Frankel and M. I. D'Angelica, "Hepatic resection for colorectal metastases," *Journal of Surgical Oncology*, vol. 109, no. 1, pp. 2–7, 2014.

[21] J. S. Tomlinson, W. R. Jarnagin, R. P. DeMatteo et al., "Actual 10-year survival after resection of colorectal liver metastases defines cure," *Journal of Clinical Oncology*, vol. 25, no. 29, pp. 4575–4580, 2007.

[22] M. C. De Jong, C. Pulitano, D. Ribero et al., "Rates and patterns of recurrence following curative intent surgery for colorectal liver metastasis: an international multi-institutional analysis of 1669 patients," *Transactions of the Meeting of the American Surgical Association*, vol. 127, pp. 84–92, 2009.

[23] D. J. Gallagher and N. Kemeny, "Metastatic colorectal cancer: from improved survival to potential cure," *Oncology*, vol. 78, pp. 237–248, 2010.

[24] B. Nordlinger, H. Sorbye, and B. Glimelius, "Perioperative FOLFOX4 chemotherapy and surgery versus surgery alone for resectable liver metastases from colorectal cancer (EORTC 40983): long-term results of a randomised, controlled, phase 3 trial," *The Lancet Oncology*, vol. 14, no. 12, pp. 1208–1215, 2013.

[25] G. Poston, R. Adam, J. Xu et al., "The role of cetuximab in converting initially unresectable colorectal cancer liver metastases for resection," *European Journal of Surgical Oncology*, vol. 43, no. 11, pp. 2001–2011, 2017.

[26] J. G. Geoghegan and J. Scheele, "Treatment of colorectal liver metastases," *The British Journal of Surgery*, vol. 86, no. 2, pp. 158–169, 1999.

[27] T. M. Drake and E. M. Harrison, "Malignant liver tumours," *Surgery*, vol. 35, no. 12, pp. 707–714, 2017.

[28] P.-A. Clavien, H. Petrowsky, M. L. DeOliveira, and R. Graf, "Strategies for safer liver surgery and partial liver transplantation," *The New England Journal of Medicine*, vol. 356, no. 15, pp. 1545–1559, 2007.

[29] P.-A. Clavien, C. E. Oberkofler, D. A. Raptis, K. Lehmann, A. Rickenbacher, and A. M. El-Badry, "What is critical for liver surgery and partial liver transplantation: Size or quality?" *Hepatology*, vol. 52, no. 2, pp. 715–729, 2010.

[30] P. Pessaux, M.-P. Chenard, P. Bachellier, and D. Jaeck, "Consequences of chemotherapy on resection of colorectal liver metastases," *Journal of Visceral Surgery*, vol. 147, no. 4, pp. e193–e201, 2010.

[31] A. W. Hemming, A. I. Reed, R. J. Howard et al., "Preoperative Portal Vein Embolization for Extended Hepatectomy," *Annals of Surgery*, vol. 237, no. 5, pp. 686–693, 2003.

[32] D. V. Sahani, M. A. Bajwa, Y. Andrabi, S. Bajpai, and J. C. Cusack, "Current status of imaging and emerging techniques to evaluate liver metastases from colorectal carcinoma," *Annals of Surgery*, vol. 259, no. 5, pp. 861–872, 2014.

[33] Y. Dong, W.-P. Wang, F. Mao, Z.-B. Ji, and B.-J. Huang, "Application of imaging fusion combining contrast-enhanced ultrasound and magnetic resonance imaging in detection of hepatic cellular carcinomas undetectable by conventional ultrasound: hepatocellular carcinoma imaging fusion," *Journal of Gastroenterology and Hepatology*, vol. 31, pp. 822–828, 2016.

[34] M. A. J. Van Den Broek, S. W. M. Olde Damink, C. H. C. Dejong et al., "Liver failure after partial hepatic resection: Definition, pathophysiology, risk factors and treatment," *Liver International*, vol. 28, no. 6, pp. 767–780, 2008.

[35] D. Ribero, G. Zimmitti, T. A. Aloia et al., "Preoperative cholangitis and future liver remnant volume determine the risk of liver failure in patients undergoing resection for hilar cholangiocarcinoma," *Journal of the American College of Surgeons*, vol. 223, no. 1, pp. 87–97, 2016.

[36] J. T. Mullen, D. Ribero, S. K. Reddy et al., "Hepatic insufficiency and mortality in 1,059 noncirrhotic patients undergoing major hepatectomy," *Journal of the American College of Surgeons*, vol. 204, no. 5, pp. 854–864, 2007.

[37] D. T. Fetzer, M. A. Rees, A. K. Dasyam, and M. E. Tublin, "Hepatic sarcoidosis in patients presenting with liver dysfunction: imaging appearance, pathological correlation and disease evolution," *European Radiology*, vol. 26, no. 9, pp. 3129–3137, 2016.

[38] M. Grąt, W. Hołõwko, Z. Lewandowski et al., "Early postoperative prediction of morbidity and mortality after a major liver resection for colorectal metastases," *HPB*, vol. 15, no. 5, pp. 352–358, 2013.

Hindawi
BioMed Research International
Volume 2018, Article ID 5830547, 7 pages
https://doi.org/10.1155/2018/5830547

Research Article

Correlation between the Expression of PD-L1 and Clinicopathological Features in Patients with Thymic Epithelial Tumors

Yanmei Chen,[1] Yuping Zhang,[2] Xiaoling Chai,[1] Jianfang Gao,[2] Guorong Chen,[3] Weifen Zhang,[4] and Yunxiang Zhang[2]

[1]*Department of Pathology, Wenzhou People's Hospital, Wenzhou, Zhejiang 325000, China*
[2]*Department of Pathology, Weifang People's Hospital, Weifang, Shandong 261041, China*
[3]*Department of Pathology, The First Affiliated Hospital, Wenzhou Medical College, Wenzhou, Zhejiang 325000, China*
[4]*School of Pharmacy, Weifang Medical University, Weifang, Shandong Province 261053, China*

Correspondence should be addressed to Yunxiang Zhang; zhangbing199592@163.com

Received 4 January 2018; Revised 7 February 2018; Accepted 13 March 2018; Published 23 April 2018

Academic Editor: Jialiang Yang

The incidence of thymic epithelial tumors (TETs) in the Chinese population was much higher than that in the North American population. In clinical treatment, the prognosis of benign tumors after surgical resection was significantly better than that of malignant tumors. Currently, the commonly used clinical indicators for TET staging included Masaoka staging and WHO (2015) pathological criteria; however, the distinction between the benign and malignant tumors and diagnosis is yet to be explored. The current study demonstrated that the expression of PD-L1 in tumor cells was correlated with the degree of TET malignancy. The quantitative analysis of PD-L1 expression in 70 cases of TET tumor samples revealed that the positive rate of PD-L1 expression in types A, AB, B1, and B2 of thymoma (40 cases) was 37.5% (15/40), which was significantly lower than that in type B3 thymoma and thymic carcinoma (76.67%, 30 cases, 23/30) as demonstrated by chi-square test ($P < 0.05$). In addition, the two methods were analyzed for the quantitative detection of PD-L1 expression. The results from the estimation of transcriptional RNA expression and quantitative protein immunohistochemistry were consistent ($r = 0.745$). Furthermore, we also analyzed PD-L1 expression level in different types of TETs from TCGA database and observed that higher PD-L1 expression was in thymic carcinoma than in thymoma. Therefore, it could be concluded that PD-L1 expression in TET cells was correlated with the degree of malignancy, whereas the estimation of PD-L1 expression was potentially applicable in the clinical staging of TETs.

1. Introduction

Thymic epithelial tumors (TETs) referred to tumors originating from thymic epithelial cells or those differentiating into the thymic epithelium, including thymoma and thymic carcinoma [1]. The incidence of thymoma in China was about 3.93/1000000, which was higher than that in the North American population (2.14/1000000). The current clinical diagnosis of benign or malignant thymic tumors was primarily carried out via the Masaoka staging and WHO (2015) pathological evaluation system. However, some of the tumors in an early stage, which was based on this classification system, still showed a relatively strong invasion [2], and, hence, surgical resection was the major therapy for such tumors. However, TETs cannot be radically cured by surgical resection for invasive thymoma, and these tumors often relapse after surgery with poor prognosis. Currently, immunotherapy based on the blockage of programmed death 1 (PD-1)/programmed death ligand 1 (PD-L1) is satisfactory in a variety of aggressive tumor species. Moreover, PD-L1 inhibited the activation of T-cells by binding to the receptor PD-1 on the surface of cytotoxic T-cells, whereby some tumors were evaded from the immune system-mediated killing through the PD-1/PD-L1 signaling pathway [3], which has been validated in animal model experiments [4]. Thus, the purpose of treating tumors can be achieved by blocking the PD-1/PD-L1 signaling

pathway to induce apoptosis. The PD-1/PD-L1 pathways have been studied in a variety of tumors, such as melanoma, ovarian cancer, colon cancer, lung cancer, and kidney cancer. The expression of PD-1 and PD-L1 may be associated with clinicopathological features and poor prognosis of malignant tumors [5–8]. Clinically, the PD-1 monoclonal antibodies, Pembrolizumab and Nivolumab, have been slightly effective in the treatment of tumors such as non-small-cell lung cancer and malignant melanoma. Thymus was an immune organ for the differentiation, development, and maturation of T-cells. PD-L1 was expressed in the thymic epithelial cells in normal thymus [9, 10]. However, only a few studies concerning the PD-1/PD-L1 in TETs are yet available globally. Thus, we aimed to explore the relationship between the expression of PD-L1/PD-1 in TETs and the correlation with the clinicopathological features. Thus, we attempted to provide a novel insight into anti-PD-1/PD-L1 treatment.

2. Materials and Methods

2.1. Clinical Information. A total of 70 cases of pathologically diagnosed TETs by surgical resection or needle biopsy were collected from 2012 to 2017 at the Department of Pathology in Weifang People's Hospital and Wenzhou People's Hospital, China. The cohort consisted of 50 thymoma and 20 thymic carcinoma cases. According to the WHO (2015) diagnostic criteria, two pathologists reviewed the section, confirmed the diagnosis, and deduced that there were sufficient tumor cells for subsequent immunohistochemistry (IHC) and genetic testing. Also, the clinicopathological data, including age, gender, with/without myasthenia gravis, with/without preoperative and postoperative radiotherapy and chemotherapy, tumor size, pathological classification, Masaoka staging, with/without lymph node, and distant metastasis were collected.

2.2. Tissue Microarray Preparation, IHC Staining, and Interpretation of PD-L1 and PD-1. A 4 × 4 array of acceptor wax blocks was made using a tissue chip instrument. The diameter of each micropore was 0.9 mm, the interval about 1 mm, and the depth about 3 mm. The IHC staining was performed on a Roche BenchMark XT fully automated IHC instrument. PD-L1 (SP142, Ventana Medical System, Tucson, AZ, USA) and PD-1 (NAT, Ventana Medical System, Tucson, AZ, USA) staining was performed according to the manufacturers' instructions. Phosphate buffer saline (PBS) was used as a negative control instead of the primary antibody. The placental tissue and tonsil tissue were used as positive control for PD-L1 and PD-1, respectively. PD-L1 was primarily located in the cell membrane; the positive color was manifested as yellow to brown linear staining on the tumor cell membrane. PD-1 was localized in the cytoplasm of the interstitial lymphocytes. The interpretation of PD-L1 results was divided into two parts: the percentage of positive cells (A) and staining intensity (B). The percentage of positive cells (A): enumeration of the percentage of stained tumor cells (0–100%), percentage of positive cells = number of positive cells/total cells × 100%. The standard of staining intensity score for positive cells: 0 point, negative; 1 point, weakly positive (light yellow); 2 points, moderately positive (brownish yellow); 3 points, strongly positive (tan). The expression of PD-L1 was evaluated by the product of positive cell percentage and the staining intensity, and the positive expression was defined when the above formula-based estimation was ≥3% [11]. The interstitial lymphocyte PD-1 in the thymoma tissues was interpreted independently. A positive result was deemed as the percentage of positive cells ≥ 5%, while the negative result was defined as <5% [12].

2.3. Detection and Interpretation of PD-L1 mRNA in RT-PCR. The wax blocks containing corresponding tumor tissues were serially sectioned into 5–8 slices, each with a thickness of 5 μm. Firstly, the total tissue RNA was extracted from the FFPE samples by RNA isolation and extraction kit (Tiangen Biochemical Beijing Co., Ltd.), followed by reverse transcription into cDNA and qPCR. The qPCR reaction system consisted of 2 μL cDNA template, 1 μL PD-L1 primers and probes (Life Technology Company, USA), 7 μL ribozyme-free water, and 10 μL Master Mix. The specific reaction program was as follows: Uracil-DNA Glycosylase incubation at 50°C for 2 min, polymerase activation at 95°C for 10 min, and PCR reaction at 95°C for 15 s and 60°C for 1 min for a total of 40 cycles. The gene mRNA expression level was detected based on the difference of Ct values (ΔCt) of the patient's target gene and the housekeeping gene as compared to the corresponding ΔCt database (the values in the database were fitted to a normal distribution curve; the expression was the relative position of the data on the curve), and the normal distribution frequency of the patient's ΔCt in the population was obtained. The relative expression level of the genes in this study was defined as follows: <0.25 was low expression; 0.25–0.75 was moderate expression; and >0.75 was high expression.

2.4. Detection of PD-L1 Expression Level in TETs from TCGA Database. 105 TET cases were selected for analysis. All cases had detailed clinical information and RNA-seq results. According to the clinical information of TET patients, all patients were divided into two categories, including thymic carcinoma and thymoma. Then we analyzed PD-L1 expression level of patients with thymic carcinoma or thymoma, respectively, and the difference of PD-L1 expression level in different categories of TETs was calculated by test. P values ≤ 0.05 were considered to be statistically significant.

2.5. Statistical Analysis. Statistical analysis was performed using SPSS 21.0 (IBM Co., Armonk, NY, USA). The comparisons between the *PD-L1* mRNA expression and PD-L1 protein and between PD-L1 protein expression and clinicopathological features were conducted by chi-square or Fisher's exact test, as appropriate. The comparisons between the PD-L1 protein and PD-1 protein expression were conducted by chi-square or Fisher's exact test, as appropriate. The correlation between the variables was analyzed by Spearman correlation analysis. Two-sided P values ≤ 0.05 were considered to be statistically significant.

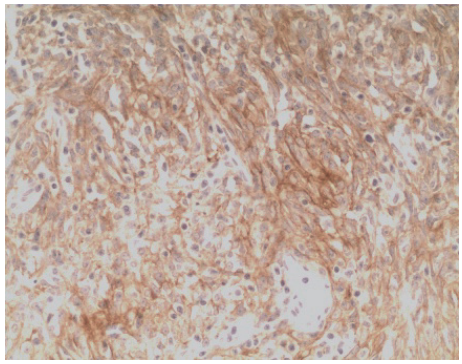


Figure 1: Type A thymoma: PD-L1-positive IHC staining, the tumor cell membrane displayed linear staining from brownish yellow to tan (Envision method, ×200).

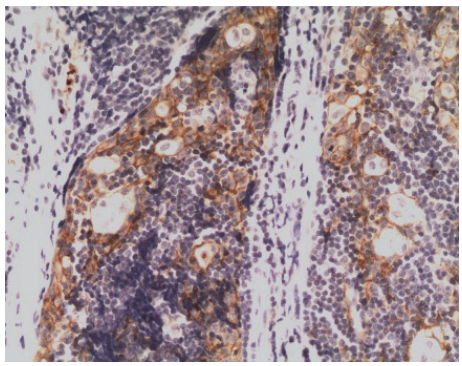


Figure 2: Type B2 thymoma: PD-L1-positive IHC staining, the tumor cell membrane displayed linear staining from brownish yellow to tan (Envision method, ×200).

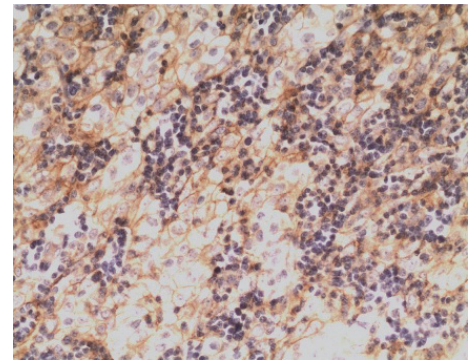


Figure 3: Type B3 thymoma: PD-L1-positive IHC staining, the tumor cell membrane displayed linear staining from brownish yellow to tan (Envision method, ×200).

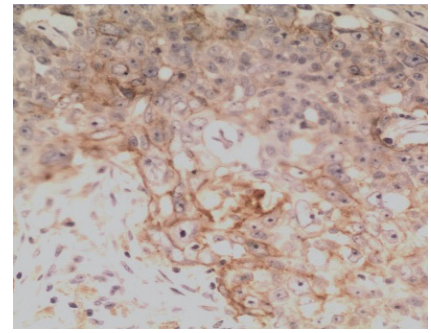


Figure 4: Thymic carcinoma (squamous cell carcinoma): PD-L1-positive IHC staining, the tumor cell membrane displayed linear staining from brownish yellow to tan (Envision method, ×200).

3. Results

3.1. Clinicopathological Characteristics of Patients. A total of 70 TET patients (30 males and 40 females) aged 29–77 (mean, 56.7 ± 12) years were included in the study. The diameters of the primary tumors were 1.2–11.74 cm. All the cases were classified according to the World Health Organization (WHO) (2015) histological criteria of thymic tumor. The cohort consisted of 11, 13, 9, 7, 10, and 20 cases of type A, type AB, type B1, type B2, type B3, and thymic carcinoma, respectively. All the thymic carcinoma cases were squamous cell carcinomas. According to the Masaoka clinical staging of thymoma, the cases of stages I, II, III, and IV were 32, 4, 20, and 14, respectively. Eleven patients exhibited comorbidity with myasthenia gravis, 37 patients underwent radiotherapy after surgery or biopsy diagnosis, and 21 patients were given chemotherapy after surgery or biopsy diagnosis; cervical lymph node metastasis, classified as type B3, was detected in 1 case of thymoma, and lymph node or other organ metastasis was observed in 13 thymic carcinoma patients.

3.2. Expression of PD-L1/PD-1 Protein in TETs. PD-L1 protein was expressed on the cell membrane of tumor cells that appeared as light yellow, brownish yellow, and tan depending on the expression level (Figures 1–4). PD-1 protein was expressed in the cytoplasm of interstitial lymphocytes (Figures 5 and 6). The positive rate of PD-L1 expression was 54.29% (38/70) in 70 cases of TETs. Among these, the positive rate of PD-L1 protein was 48% (24/50) and 70% (14/20) in thymoma and thymic carcinoma tissues, respectively; however, there was no significant difference in the positive rates between the two groups ($\chi^2 = 2.786$, $P > 0.05$). Considering that the biological behaviors of type B3 thymoma were similar to that of thymic carcinoma, it was classified as thymus carcinoma for analysis. The positive rate of PD-L1 in type B3 thymoma and thymic carcinoma (76.67%, 23/30) was significantly higher than that in other types of thymomas (37.5%, 15/40) ($\chi^2 = 10.597$, $P < 0.05$). In 20 cases of thymic carcinoma tissues, the rate of PD-1 was positive in 13 cases of tumor-infiltrating lymphocytes (TILs) with a positive rate of 65%, and in 14 cases of tumor cells with a positive rate of 70%. In addition, PD-1 was positively correlated with PD-L1 expression in tumor cells ($P < 0.05$, correlation coefficient $r = 0.663$).

3.3. Relationship between PD-L1 Protein Expression and Clinicopathological Features in TETs. The PD-L1 protein expression in TETs was not associated with gender, age, tumor size, with/without metastasis, and with/without myasthenia gravis

Table 1: Relationship between the expression of PD-L1 protein and clinicopathological features in TETs.

Clinicopathological factors	Case number	PD-L1 expression		Positive rate (%)	χ^2	P value
		+	−			
Gender						
Male	30	20	10	33.33	3.243	0.092
Female	40	18	22	45.00		
Age						
≤50 years	16	7	9	43.75	0.928	0.399
>50 years	54	31	23	57.41		
Tumor size						
≤4 cm	41	22	19	53.66	0.016	1.000
>4 cm	29	16	13	55.17		
Histological classification						
Type A	11	4	7	36.36	20.648	0.001*
Type AB	13	2	11	15.39		
Type B1	9	3	6	33.33		
Type B2	7	6	1	85.71		
Type B3	10	9	1	90.00		
Type C thymic carcinoma	20	14	6	70.00		
Masaoka-Koga staging						
Stage I	32	10	22	31.25	12.402	0.004*
Stage II	4	3	1	75.00		
Stage III	20	15	5	75.00		
Stage IV	14	10	4	71.42		
With/without myasthenia gravis						
Yes	11	7	4	63.64	0.460	0.533
No	59	31	28	52.54		
Metastasis						
Yes	14	10	4	71.43	2.072	0.231
No	56	28	28	50.00		
Radiotherapy						
Yes	37	26	11	70.27	8.081	0.008*
No	33	12	21	36.36		
Chemotherapy						
Yes	21	16	5	76.19	5.800	0.02*
No	49	22	27	44.90		

* $P < 0.05$.

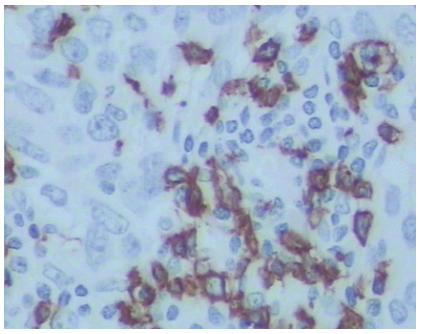


Figure 5: Thymic carcinoma: PD-l-positive IHC staining, the cytoplasmic staining of the interstitial immune cells was positive, and the tumor cells staining was negative (Envision method, ×400).

symptoms; however, it was correlated with WHO histological classification, Masaoka-Koga staging, radiotherapy, and chemotherapy. The difference was statistically significant (all $P < 0.05$) (Table 1).

3.4. Quantitation of PD-L1 mRNA Expression. Of the 70 cases of TETs, the *PD-L1* mRNA was highly expressed in 17 cases, with positive PD-L1 protein detection and positive coincidence rate of 100%. On the other hand, *PD-L1* mRNA was moderately expressed in 22 cases, of which the PD-L1 protein was positive in 15 cases with a positive coincidence rate of 68.18%. The *PD-L1* mRNA was lowly expressed in 17 cases, in which the PD-L1 protein was positive in only 1 case, with a negative coincidence rate of 94.12%. The concentration and

Table 2: Correlation between mRNA and protein expression of PD-L1.

PD-L1 mRNA expression	PD-L1 protein expression		Coincidence rate (%)	χ^2	*P* value
	+	−			
High expression	17	0	100		
Moderate expression	15	7	68.18	32.392	<0.001
Low expression	1	16	94.12		

≤Spearman correlation coefficient $r = 0.745$.

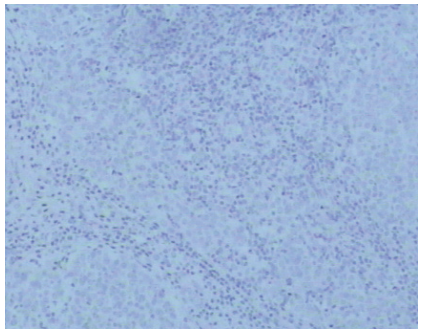


Figure 6: Type C thymoma: PD-1-negative IHC staining (Envision method, ×100).

purity of RNA extracted from 14 samples were infinitesimally low, and, hence, these were regarded as unqualified samples and not included in the statistical analysis. Five cases of PD-L1 protein were detected as positive. The coincidence rate of the two detection methods was 85.71%. Spearman correlation analysis showed that *PD-L1* mRNA expression was positively correlated with PD-L1 protein expression ($P < 0.001$, correlation coefficient = 0.745, Table 2).

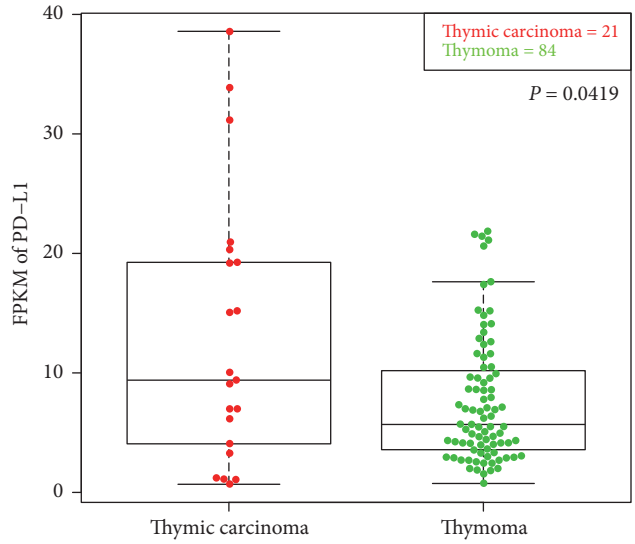


Figure 7: Analyzing PD-L1 expression in different types of TETs from TCGA database. We compared PD-L1 expression in thymoma with that in thymic carcinoma, and the difference was significant ($P < 0.05$).

3.5. Analyzing PD-L1 Expression in Different Types of TETs from TCGA Database. We obtained 105 TETs cases with detailed clinical information and RNA-seq results from TCGA database. Based on the clinical characteristics, all TET patients ($n = 105$) could be simply divided into two categories, including thymoma ($n = 84$) and thymic carcinoma ($n = 21$). We found that the median PD-L1 expression in patients with thymoma was 5.68, and the median PD-L1 expression in patients with thymic carcinoma was 9.39 (Figure 7). The difference was significant ($P = 0.0419$). The result further suggests that high PD-L1 expression is substantially correlated with malignancy of TET tumors.

4. Discussion

The study of the PD-1/PD-L1 pathway in many tumors has been under intensive focus. However, only a few reports about PD-1/PD-L1 in TETs are available in recent years. In this study, we analyzed the expression of PD-L1 protein in 70 cases of TETs. The results showed that the positive expression rate of PD-L1 protein in TETs was 54.29%. Previous studies [10, 11, 13–18] demonstrated that the positive expression rate of PD-L1 protein in thymoma was 18–92%. The results of PD-L1 protein quantitation proposed that the differences may be attributed to the following reasons: (1) the clone number of the PD-L1 antibody varied in different studies. For example, the staining model of PD-L1 antibody (SP142) used in this experiment was from weak to strong staining of the cell membrane, while the PD-L1 antibody with clone numbers 5H1 and 15 provided a diffused and consistent staining model, respectively [10, 13]. Thus, the SP142 staining model of PD-L1 antibody may be optimal for the quantitative assessment, with high reliability. (2) The standards for the interpretation of positive expression were inconsistent. For example, some PD-L1 clone numbers stained not only the tumor cells but also interstitial cells. The two kinds of positive cells were included during interpretation, which led to an increase in the positive rate of TETs [15]. (3) The subjectivity of the scoring was noted in positive interpretation by different observers, and the repeatability was poor. (4) Sample size and sample selection errors, for example, high percentage of type A and type AB thymoma in the cohort, would lead to a low TETs positive rate [13]. (5) The heterogeneity of PD-L1 expression in different sampling sites was associated with the differences.

Moreover, this study, for the first time, used real-time fluorescence qPCR to detect the *PD-L1* mRNA expression in TETs, which also provided a credible basis for the estimation of PD-L1 protein. The results showed that the expression

of *PD-L1* mRNA in high- and low-expression groups was in agreement with that by IHC detection; however, the difference primarily occurred in the moderate-expression group, in which the concordance rate with protein was 68.18%. These differential phenomena might be attributed to the following: real-time fluorescence qPCR detected the mRNA expression level in all cells, including tumor and interstitial cells; however, the IHC interpretation of PD-L1 protein was confined to the tumor cells, and the analysis did not include the nontumor lymphocytes that expressed the PD-L1 protein. Furthermore, TETs were different from other solid tumors, especially type B1 and B2 thymoma. The TETs were similar to the normal thymus, with relatively abundant nontumor immune cells. A total of 14 cases of patients with tissues expressing *PD-L1* mRNA did not reveal any unqualified RNA. However, IHC showed positive PD-L1 expression in 5 cases. Although the sensitivity of mRNA detection was higher than that of IHC, it could not clearly identify the same in tumor cells, and the specificity was poor. On the other hand, the RNA was degraded due to the method of sample fixation, the length of time, and the long shelf-life, which led to failed obtaining results. Although a certain degree of subjectivity was noted in the selection of IHC antibody and the judgment of positive intensity and positive threshold, the percentage of false-positive and false-negative results would appear with advantages of simple operation, mature technology, low cost, and improvement in the drawbacks by quality control. Therefore, the IHC method can be used as a detection indicator for PD-L1 in TETs with its unique advantages of simple operation and cost-efficiency.

Studies have found that the positive expression of the PD-L1 protein was an adverse prognostic factor [19, 20]. The high expression of PD-L1 in thymoma can be an independent risk factor for tumor recurrence and predict a poor overall survival [10, 13]. The prognosis of cancer patients was closely related to the clinicopathological features. WHO classification, Masaoka staging, and preoperative treatment remarkably affected the overall survival rate of TETs [11]. However, there are still some paradoxical observations in previous reports [21, 22]. In order to further verify the correlation between the expression of PD-L1 protein and the biological behaviors of TETs, the present study analyzed the relationship between the expression of PD-L1 protein and clinicopathological features of TETs. The results showed that the positive expression of PD-L1 protein was correlated with the WHO histological classification and Masaoka-Koga staging of TETs. The positive expression rate of PD-L1 protein in type B3 thymoma and thymic carcinoma was significantly higher than that in other thymoma subtypes. Furthermore, we reconfirmed our observations by analyzing TCGA data. These results indicated that the expression of PD-L1 protein was correlated with the invasive and malignant degree of the tumors.

A majority of the TETs were composed of a mixture of neoplastic epithelial and nonneoplastic lymphocytes. Moreover, the detection of PD-L1 alone cannot comprehensively determine the prognosis of TETs. PD-1 is primarily expressed on activated T lymphocytes. Studies have shown that an increase of PD-1-positive TILs suggested a poor prognosis [23, 24]. Thymic cancer differed from that of thymoma in both prognosis and molecular phenotype. Based on these above speculations, the relationship between PD-L1 and PD-1 in thymic carcinoma was further evaluated in this study. The results showed that the positive expression rate of TILs PD-1 was 65% in 20 cases of thymic carcinoma. After further analysis of the correlation between PD-1 and PD-L1, we found that the expression of PD-1 and PD-L1 in thymic carcinoma was positively correlated, thereby indicating that TILs PD-1 together with tumor cell PD-L1 can reflect the activation state of the PD-1/PD-L1 pathway. This experiment, for the first time, proved a positive correlation between PD-1 and PD-L1 in thymic carcinoma. Immunotherapy targeting PD-1/PD-L1 exerts antitumor effects in several malignant tumors, and we expected this to be demonstrated in thymic cancers, especially in patients with thymic carcinoma insensitive to chemotherapy. Previous studies showed that PD-L1-positive patients responded better to immune checkpoint inhibitors than the PD-L1-negative patients in other malignant tumors [25, 26]. Based on the above findings and the high expression of PD-L1 protein in thymic carcinoma in the current study, we speculated that the thymic carcinoma with PD-1/PD-L1 could be treated as the target. Nevertheless, this study also had some limitations. The number of samples collected was relatively small due to the rarity of thymic carcinoma. Hence, further prospective studies with large sample size are essential to verify these findings.

The present study also found that PD-L1 protein expression was associated with radiotherapy and chemotherapy; however, the underlying mechanism was yet unknown. Katsuya et al. [17] reported that the expression of PD-L1 in tumor cells and PD-1 in interstitial lymphocytes was increased significantly in TETs undergoing chemotherapy. This result was beneficial for postchemotherapy immunotherapy and proposed the putative combination of chemotherapy and immunotherapy in treating TETs.

5. Conclusion

Overall, the current findings indicated that the expressions of PD-L1 protein and mRNA differed in thymoma and thymic carcinomas, and PD-L1 may serve as a potential marker of invasiveness and prognosis. In addition, the high expression of PD-L1 and PD-1 in TETs has also made it possible to clinically adopt the immunotherapy for targeting PD-1/PD-L1.

Conflicts of Interest

The authors declare no conflicts of interest.

Authors' Contributions

Yanmei Chen and Yuping Zhang contributed equally to this study.

Acknowledgments

This study was funded by the Key Science and Technology Program of Shandong Province (2015GSF118168), the

Weifang Development Project (20100220), and National Natural Science Foundation of China (mo. 81774125). The authors also acknowledge the Genesis Beijing Co., Ltd. (Beijing 100102, China).

References

[1] F. Detterbeck and A. M. Parsons, "Thymic tumor: a review of current diagnosis, classification, and treatment," in *Thoracic and Esophageal Surgery*, G. A. Patterson, J. D. Cooper, J. Deslauriers et al., Eds., pp. 1589–1614, Elsevier, Philadelphia, Pennsylvania, 3rd edition, 2008.

[2] F. C. Detterbeck and A. M. Parsons, "Management of stage I and II thymoma," *Thoracic Surgery Clinics*, vol. 21, no. 1, pp. 59–67, 2011.

[3] L. Han, R. Li, X. Chen et al., "The expression and clinical significance of B7 family molecules and their receptor PD-1 in human NK/T-cell lymphoma," *Chinese Journal of Clinical Oncology*, vol. 41, no. 6, pp. 363–367, 2014.

[4] C. S. Lages, I. Lewkowich, A. Sproles, M. Wills-Karp, and C. Chougnet, "Partial restoration of T-cell function in aged mice by in vitro blockade of the PD-1/PD-L1 pathway," *Aging Cell*, vol. 9, no. 5, pp. 785–798, 2010.

[5] R. H. Thompson, S. M. Kuntz, B. C. Leibovich et al., "Tumor B7-H1 is associated with poor prognosis in renal cell carcinoma patients with long-term follow-up," *Cancer Research*, vol. 66, no. 7, pp. 3381–3385, 2006.

[6] H. Dong, S. E. Strome, and D. R. Salomao, "Tumor-associated B7-H1 promotes T-cell apoptosis: a potential mechanism of immune evasion," *Nature Medicine*, vol. 8, no. 8, pp. 793–800, 2002.

[7] Y.-B. Chen, C.-Y. Mu, and J.-A. Huang, "Clinical significance of programmed death-1 ligand-1 expression in patients with non-small cell lung cancer: A 5-year-follow-up study," *Tumori*, vol. 98, no. 6, pp. 751–755, 2012.

[8] V. Velcheti, K. A. Schalper, D. E. Carvajal et al., "Programmed death ligand-1 expression in non-small cell lung cancer," *Laboratory Investigation*, vol. 94, no. 1, pp. 107–116, 2014.

[9] J. A. Brown, D. M. Dorfman, F.-R. Ma et al., "Blockade of programmed death-1 ligands on dendritic cells enhances T cell activation and cytokine production," *The Journal of Immunology*, vol. 170, no. 3, pp. 1257–1266, 2003.

[10] S. K. Padda, J. W. Riess, E. J. Schwartz et al., "Diffuse high intensity PD-L1 staining in thymic epithelial tumors," *Journal of Thoracic Oncology*, vol. 10, no. 3, pp. 500–508, 2015.

[11] Y. Katsuya, Y. Fujita, H. Horinouchi, Y. Ohe, S. Watanabe, and K. Tsuta, "Immunohistochemical status of PD-L1 in thymoma and thymic carcinoma," *Lung Cancer*, vol. 88, no. 2, pp. 154–159, 2015.

[12] J. M. Taube, A. Klein, J. R. Brahmer et al., "Association of PD-1, PD-1 ligands, and other features of the tumor immune microenvironment with response to anti-PD-1 therapy," *Clinical Cancer Research*, vol. 20, no. 19, pp. 5064–5074, 2014.

[13] S. Yokoyama, H. Miyoshi, T. Nishi et al., "Clinicopathologic and Prognostic Implications of Programmed Death Ligand 1 Expression in Thymoma Presented at the Sixteenth World Conference on Lung Cancer, Denver, CO, September 6-9, 2015.," *The Annals of Thoracic Surgery*, vol. 101, no. 4, pp. 1361–1369, 2016.

[14] M. Tiseo, A. Damato, L. Longo et al., "Analysis of a panel of druggable gene mutations and of ALK and PD-L1 expression in a series of thymic epithelial tumors (TETs)," *Lung Cancer*, vol. 104, pp. 24–30, 2017.

[15] A. M. Marchevsky and A. E. Walts, "PD-L1, PD-1, CD4, and CD8 expression in neoplastic and nonneoplastic thymus," *Human Pathology*, vol. 60, pp. 16–23, 2017.

[16] S. Yokoyama, H. Miyoshi, K. Nakashima et al., "Prognostic value of programmed death ligand 1 and programmed death 1 expression in thymic carcinoma," *Clinical Cancer Research*, vol. 22, no. 18, pp. 4727–4734, 2016.

[17] Y. Katsuya, H. Horinouchi, T. Asao et al., "Expression of programmed death 1 (PD-1) and its ligand (PD-L1) in thymic epithelial tumors: Impact on treatment efficacy and alteration in expression after chemotherapy," *Lung Cancer*, vol. 99, pp. 4–10, 2016.

[18] A. Weissferdt, J. Fujimoto, N. Kalhor et al., "Expression of PD-1 and PD-L1 in thymic epithelial neoplasms," *Modern Pathology*, vol. 30, no. 6, pp. 826–833, 2017.

[19] C. Wu, Y. Zhu, J. Jiang, J. Zhao, X.-G. Zhang, and N. Xu, "Immunohistochemical localization of programmed death-1 ligand-1 (PD-L1) in gastric carcinoma and its clinical significance," *Acta Histochemica*, vol. 108, no. 1, pp. 19–24, 2006.

[20] J. Hamanishi, M. Mandai, M. Iwasaki et al., "Programmed cell death 1 ligand 1 and tumor-infiltrating $CD8^+$ T lymphocytes are prognostic factors of human ovarian cancer," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 104, no. 9, pp. 3360–3365, 2007.

[21] S. K. Padda, J. W. Riess, E. J. Schwartz et al., "Diffuse High Intensity PD–L1 Staining in Thymic Epithelial Tumors," *Journal of Thoracic Oncology*, vol. 10, no. 3, pp. 500–508, 2015.

[22] K. C. Arbour et al., "Expression of PD-L1 and other immunotherapeutic targets in thymic epithelial tumors," *PLoS One*, vol. 12, no. 8, article no. e0182665, 2017.

[23] S. Muenst, S. D. Soysal, F. Gao, E. C. Obermann, D. Oertli, and W. E. Gillanders, "The presence of programmed death 1 (PD-1)-positive tumor-infiltrating lymphocytes is associated with poor prognosis in human breast cancer," *Breast Cancer Research and Treatment*, vol. 139, no. 3, pp. 667–676, 2013.

[24] R. H. Thompson, H. Dong, C. M. Lohse et al., "PD-1 is expressed by tumor-infiltrating immune cells and is associated with poor outcome for patients with renal cell carcinoma," *Clinical Cancer Research*, vol. 13, no. 6, pp. 1757–1761, 2007.

[25] G. K. Philips and M. Atkins, "Therapeutic uses of anti-PD-1 and anti-PD-L1 antibodies," *International Immunology*, vol. 27, no. 1, pp. 39–46, 2015.

[26] X. Meng, Z. Huang, F. Teng, L. Xing, and J. Yu, "Predictive biomarkers in PD-1/PD-L1 checkpoint blockade immunotherapy," *Cancer Treatment Reviews*, vol. 41, no. 10, pp. 868–876, 2015.

Hindawi
BioMed Research International
Volume 2018, Article ID 4761064, 10 pages
https://doi.org/10.1155/2018/4761064

Research Article

Identification of Key Genes and miRNAs in Osteosarcoma Patients with Chemoresistance by Bioinformatics Analysis

Binbin Xie,[1] Yiran Li,[1] Rongjie Zhao,[1] Yuzi Xu,[2] Yuhui Wu,[1] Ji Wang,[3,4] Dongdong Xia,[5] Weidong Han,[1] and Dake Chen[6]

[1]*Department of Medical Oncology, Sir Run Run Shaw Hospital, School of Medicine, Zhejiang University, Hangzhou, China*
[2]*Department of Sports Medicine, School of Medicine, Zhejiang University, Zhejiang, China*
[3]*Department of Surgical Oncology, Sir Run Run Shaw Hospital, College of Medicine, Zhejiang University, Hangzhou, Zhejiang 310016, China*
[4]*Biomedical Research Center and Key Laboratory of Biotherapy of Zhejiang Province, Hangzhou, Zhejiang 310016, China*
[5]*Orthopedic Department, Ningbo First Hospital, Ningbo 315000, China*
[6]*Department of Urology, Wenzhou People's Hospital, Wenzhou, Zhejiang, China*

Correspondence should be addressed to Dongdong Xia; xiadongdong@qq.com, Weidong Han; hanwd@zju.edu.cn, and Dake Chen; 61502485@qq.com

Received 23 November 2017; Revised 21 February 2018; Accepted 4 March 2018; Published 22 April 2018

Academic Editor: Jialiang Yang

Chemoresistance is a significant factor associated with poor outcomes of osteosarcoma patients. The present study aims to identify Chemoresistance-regulated gene signatures and microRNAs (miRNAs) in Gene Expression Omnibus (GEO) database. The results of Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) included positive regulation of transcription, DNA-templated, tryptophan metabolism, and the like. Then differentially expressed genes (DEGs) were uploaded to Search Tool for the Retrieval of Interacting Genes (STRING) to construct protein-protein interaction (PPI) networks, and 9 hub genes were screened, such as fucosyltransferase 3 (Lewis blood group) (FUT3) whose expression in chemoresistant samples was high, but with a better prognosis in osteosarcoma patients. Furthermore, the connection between DEGs and differentially expressed miRNAs (DEMs) was explored. GEO2R was utilized to screen out DEGs and DEMs. A total of 668 DEGs and 5 DEMs were extracted from GSE7437 and GSE30934 differentiating samples of poor and good chemotherapy reaction patients. The Database for Annotation, Visualization, and Integrated Discovery (DAVID) was used to perform GO and KEGG pathway enrichment analysis to identify potential pathways and functional annotations linked with osteosarcoma chemoresistance. The present study may provide a deeper understanding about regulatory genes of osteosarcoma chemoresistance and identify potential therapeutic targets for osteosarcoma.

1. Introduction

Osteosarcoma is one of the most common primary malignant bone tumors in children and adolescents. The worldwide morbidity rates of osteosarcoma are approximately with an average incidence of 3.1 per million for each stage and 4.4 per million for groups <25 years old. Additionally, there is a bimodal age distribution: individuals aged 25–60 years and elderly individuals, respectively. In America the Annual age-standardized incidence of osteosarcoma has reached stabilization from 1976 to 2005 [1–5].

Many factors are associated with tumor genesis, including high birth weight [6], pubertal hormones [7], and germline genetic variants [8, 9]. The common subtypes are osteoblastic, chondroblastic, and fibroblastic osteosarcomas, which may account for 70–80% of total cases [10]. The standard therapy consists of neoadjuvant chemotherapy (NACT), surgical removal of the primary tumor, and adjuvant chemotherapy. Before the 1970s, no more than 20 percent of patients were alive after 5 years when excision was major therapeutic measure for osteosarcoma [11, 12], while it increased to 60–70 percent for children and young adults with localized

disease after the chemotherapy was used as adjuvant therapy for surgical resection [13, 14]. The current predicament of osteosarcoma treatment is the five-year survival rate does not exceed 25% for patients aged 2–68 with poor initial response tending to have adverse outcomes [15]. Therefore, to improve and modify chemotherapy regimens, an increasing number of pharmacogenomics studies on osteosarcoma have been going on for some time, such as drug reactions and toxicity.

Multidrug resistance protein 1 (MDR1) that is encoded by gene ATP-binding cassette, subfamily B (MDR/TAP), member 1 (ABCB1) has been shown to serve as a plausible factor in doxorubicin resistance, which was validated to be linked with poor outcomes in many osteosarcoma studies [16, 17], but whether there would be more valuable biomarkers remained to be explored. In recent years, microarray technology has substantially promoted the advance of understanding the mechanisms underlying diseases. Additionally, the rapid development of bioinformatics enables us to comprehensively screen out the hub genes associated with chemoresistance by the process of high-throughput microarrays. MicroRNAs (miRNAs), a group of highly conserved short noncoding small RNAs including generally 18–25 nucleotides in length, can suppress the translation of mRNA and cleave it by the modality of base-pairing to the target genes' $3'$ untranslated region [18–20].

In the present study, we analyzed the data of GSE87437 and GSE30934 submitted by Serra Mand and Kobayashi E. et al., respectively, to get 668 differentially expressed genes (DEGs) and 5 differentially expressed miRNA (DEMs) between samples of poor and good chemotherapy reaction patients in GEO2R (http://www.ncbi.nlm.nih.gov/geo/geo2r/). To further understand the function of genes, Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes pathway (KEGG), Protein-Protein Interaction (PPI) networks, and the connections among DEGs and DEMs were performed in sequence. We selected chemoresistance development related key genes and provided theoretical foundations for modifying and improving osteosarcoma treatment methods.

2. Materials and Methods

2.1. Microarray Data. The dataset of GSE87437, gene expression array, and GSE30934, miRNA expression array, included 10 and 8 samples from poor chemotherapy reaction patients and 11 and 16 samples from good ones, respectively. Moreover, the former dataset was based on GPL570 platform ([HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array) and the latter one was based on GPL10312 platform (3D-Gene Human miRNA Oligo chip v12-1.00).

2.2. Identification of DEGs. GEO2R, an R-associated web application, was applied to filtrate DEGs between good chemotherapy reaction samples and poor chemotherapy reaction samples. In total, 21 samples in GSE87437 and 24 samples in GSE30934 were divided into two groups, respectively, and the concrete grouping schemes were already shown in microarray data. The $P < 0.05$ and $|\log \mathrm{FC}| \geqslant 1$ were considered as cutoff criterion. All results of DEGs were downloaded in text format, hierarchical clustering analysis being conducted later in Morpheus (https://software.broadinstitute.org/morpheus/).

2.3. GO and Pathway Enrichment Analysis of DEGs. The online tool, Database for Annotation, Visualization and Integrated Discovery (DAVID, https://david.ncifcrf.gov/) provided comprehensive information for list of genes by GO and KEGG pathway analyses. In addition, GO enrichment analysis included three different aspects: biological process (BP), molecular function (MF), and cellular component (CC) [21]. KEGG enrichment analysis was associated with genomic information's functional interpretation and practical application [22]. The screened DEGs were uploaded to DAVID V6.8 to perform GO and KEGG pathway analysis with the criterion of $P < 0.05$, the results of which were downloaded in text format.

2.4. PPI Networks Construction and Module Analysis. To analyze the connection among proteins, DEGs were uploaded to Search Tool for the Retrieval of Interacting Genes (STRING, https://string-db.org/), a database covering 9,643,763 proteins from 2,031 organisms, and the result whose minimum interaction score was 0.4 was visualized in Cytoscape [23, 24]. Furthermore, the Molecular Complex Detection (MCODE) was used to screen out significant modules based on the constructed PPI networks with the criteria of degree cutoff = 2, node density cutoff = 0.1, node score cutoff = 0.2, k-core = 2, and max. depth = 100 and hub genes were exported. The functional enrichment analysis of genes in each module was performed in DAVID. Besides, the genes in each module were uploaded to DAVID and KEGG pathway enrichment analysis was conducted with the condition of $P < 0.05$.

2.5. Survival Analysis of Hub Genes. The series matrix of GSE21257 that contained osteosarcoma patients' prognostic information was downloaded from GEO database. The patients were split into two groups, high expression and low expression, according to the expression level of a specific hub gene. The data was processed by graphpad prism software and then exported the results.

2.6. Prediction of miRNA Targets. DEMs were acquired by the parallel method of DEGs mentioned above. miRWalk1.0 (http://zmf.umm.uni-heidelberg.de/apps/zmf/mirwalk/index.html), an integrated miRNA target prediction platform including 10 databases (DIANAmT, miRanda, miRDB, miRWalk, RNAhybrid, PICTAR4, PICTAR5, PITA, RNA22, and Targetscan), was utilized to explore the correlation between DEMs and DEGs. Besides, different colors were used to indicate the degrees of connections. For example, red color represented strong correlation.

3. Results

3.1. Identification of DEGs. A total of 668 DEGs were obtained from GSE87437 in the poor chemotherapy response samples and compared with those of good response with the criteria of $P < 0.05$ and $|\log \mathrm{FC}| \geqslant 1.0$, comprising 422

Table 1: Key differentially expressed genes (DEGs) obtained from GSE87374.

Gene symbol	Log FC	*P* value
ZNRD1	1.4	0.01504282
CDK1	1.31	0.03504221
MYH7B	1.19	0.03527194
GPR68	1.18	0.03724459
CAT	−1.15	0.03367398
FUT3	1.69	0.00007205
IMPG2	1.11	0.04785566
GPR180	−1.03	0.00511768
ANPEP	1.32	0.0354653

upregulated genes and 246 downregulated genes. The key DEGs are displayed in Table 1.

3.2. Hierarchical Clustering Analysis of DEGs. Hierarchical clustering analysis was conducted through Morpheus, a web-based online tool, with the series matrix data of the DEGs. The heat map is shown in Figure 1 (top 50 upregulated and 50 downregulated genes).

3.3. GO Term Enrichment Analysis. In order to understand the function of the identified DEGs deeply, GO and KEGG analyses were performed in DAVID, respectively. The result of GO analysis showed that DEGs were enriched in biological process (BP), including positive regulation of transcription and DNA-templated, positive regulation of sequence-specific DNA binding transcription factor activity, nitric oxide mediated signal transduction, positive regulation of transcription from RNA polymerase II promoter, and regulation of phosphatidylinositol 3-kinase signaling. As for molecular function (MF), the DEGs were enriched in estrogen response element binding, Rac guanyl-nucleotide exchange factor activity, calcium ion binding, zinc ion binding, and phosphatidylinositol-4,5-bisphosphate 3-kinase activity. Besides, Cellular Component (CC) analysis showed that the DEGs were enriched in proteinaceous extracellular matrix, cell surface, P granule, integral component of plasma membrane, and endocytic vesicle membrane, as shown in Figure 2.

3.4. KEGG Pathway Analysis. KEGG pathway analysis showed that DEGs were mainly involved in tryptophan metabolism, oxytocin signaling pathway, glyoxylate and dicarboxylate metabolism, cAMP signaling pathway, and dopaminergic synapse (Figure 2).

3.5. PPI Networks and Modules Selection. The PPI networks of DEGs were composed of 432 nodes and 428 edges (Figure 3). Then the networks were imported into Cytoscape software, analyzed by using plug-ins MCODE. Eventually, 3 significant modules were selected (Figure 4), and the KEGG pathway was mainly associated with ribosome biogenesis in eukaryotes, calcium signaling pathway, arachidonic acid metabolism, proteoglycans in cancer, and linoleic acid metabolism (Figure 4).

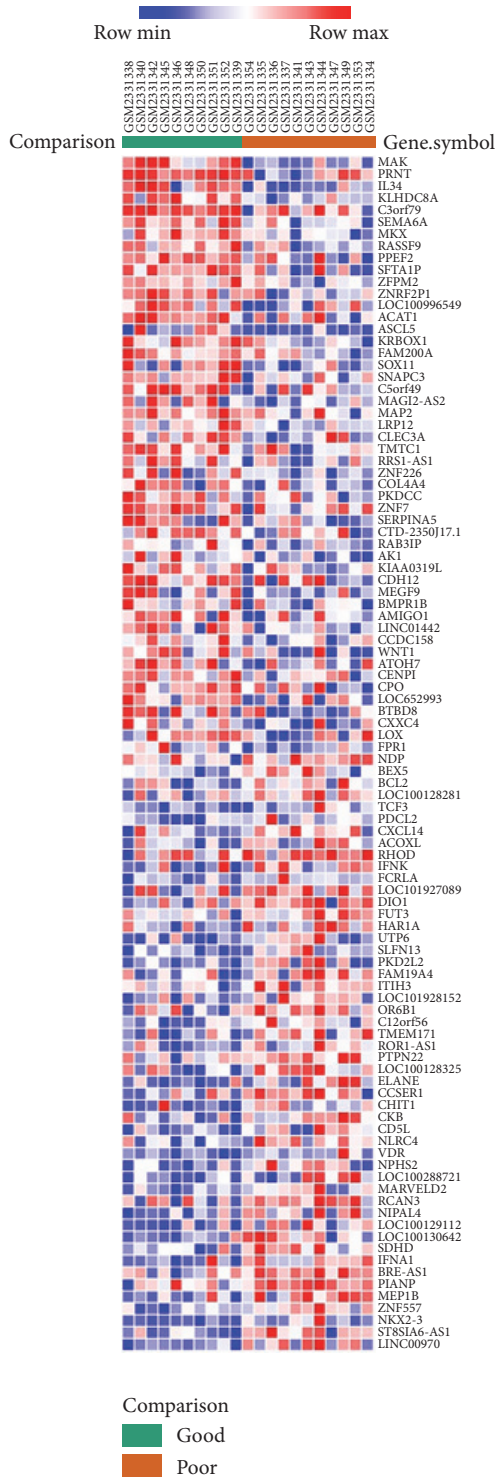


Figure 1: Heat map of the top 100 DEGs (50 upregulated genes and 50 downregulated genes). Red: up-regulation; Blue: down-regulation.

3.6. Hub Genes and Survival Analysis. 9 hub genes were screened out, including zinc ribbon domain containing 1 (ZNRD1), myosin heavy chain 7B (MYH7B), G protein-coupledreceptor 68 (GPR68), catalase (CAT), fucosyltransferase 3 (Lewis blood group) (FUT3), interphotoreceptor matrix

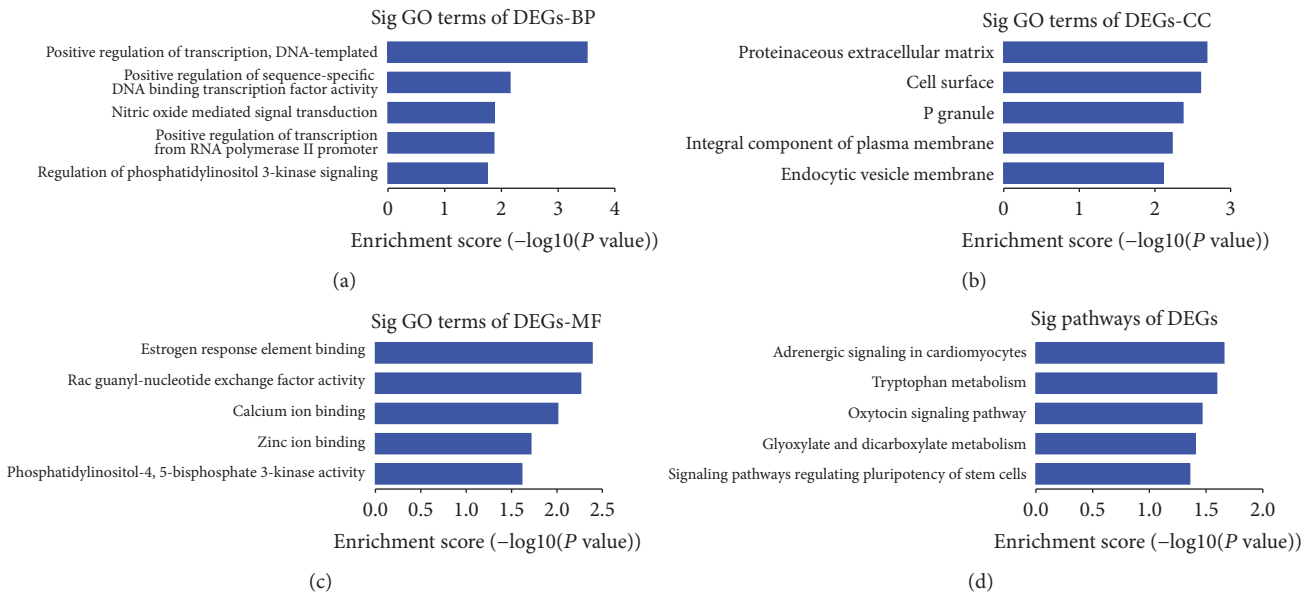


Figure 2: GO and KEGG pathway analysis of DEGs associated with osteosarcoma. (a) Top 5 significantly enriched biological processes in DEGs. (b) Top 5 significantly enriched cell component in DEGs. (c) Top 5 significantly enriched molecular function in DEGs. (d) Top 5 significantly enriched KEGG pathway in DEGs.

proteoglycan 2 (IMPG2), G protein-coupled receptor 180 (GPR180), alanyl aminopeptidase, membrane (ANPEP), and cyclin dependent kinase 1 (CDK1) (Table 1). Next, survival analysis of these genes in GSE21257 which contained patients' survival prognostic information showed that osteosarcoma patients with high mRNA expression of FUT3 meant a better overall survival (OS) despite its high expression in poor chemotherapy response samples (Figure 5). Additionally, the survival prognostic information of GPR180 and CDK1 was not included in GSE21257.

3.7. MiRNA–DEG Pairs. After the differentially expressed analysis for the data of GSE30934, a total of 5 DEMs were obtained between the poor chemotherapy response samples compared with that of good response with the criteria of $P < 0.05$ and $|\log \mathrm{FC}| \geqslant 1.0$ (Table 2). Next, basing on miRWalk1.0 database (http://zmf.umm.uni-heidelberg.de/apps/zmf/mirwalk/index.html), the relationship between miRNAs and DEGs was acquired and different kinds of colors were on behalf of the number of miRNA–DEG pairs in different database which stand for the degrees of connection. For example, red color represented to a strong correlation (Figure 6). After comparing the targets with hub genes, we found that ZNRD1 was the potential target of hsa-miR-543, while CAT was the potential target of hsa-miR-518f. Both hsa-miR-543 and hsa-miR-518f matched the regulated gene in expression trends.

Table 2: Key differentially expressed genes (DEGs) obtained from GSE30934.

miRNA_ID	Log FC	*P* value
hsa-miR-543	−3.429933	0.00192
hsa-miR-409-5p	−2.70517	0.00729
hsa-miR-518f	1.448711	0.02332
hsa-miR-154	−2.619116	0.03838
ebv-miR-BART1-3p	−1.358071	0.04733

4. Discussion

In the present study, we observed whether there were more valuable genes like ABCB1 which could help improve and modify chemotherapy regimens in osteosarcoma. To find out the specific chemotherapy response-associated DEGs, we analyzed the osteosarcoma gene expression array of GSE87437 in GEO2R, where a total number of 668 DEGs were obtained between good and poor chemotherapy response samples. Besides, to further understand the potential biological functions, we conducted GO, KEGG, and STRING analyses. Subsequently, on the foundation of PPI networks, the selection of 9 hub genes and their survival prognosis were completed. In terms of the increasingly prominent role of miRNA in cancer, DEMs of osteosarcoma miRNA expression array of GSE30934 was screened out in the same way and criteria like DEGs, and DEGs-miRNA network was constructed to show relationship between them [25].

Our results showed that many genes and miRNAs may have functions in the development of chemoresistance in osteosarcoma and have the potential to become treatment targets. Here, we exclusively focused on 9 hub genes and two miRNAs. Firstly, 9 hub genes consisted of ZNRD1, MYH7B, GPR68, CAT, FUT3, IMPG2, GPR180, ANPEP, and CDK1. Our data showed that the expression of ZNRD1 was upregulated in chemoresistance osteosarcoma samples. Previous studies demonstrated that methotrexate-resistant, vincristine-resistant, multidrug resistant phenotypes of gastric cancer cells could be regulated by the

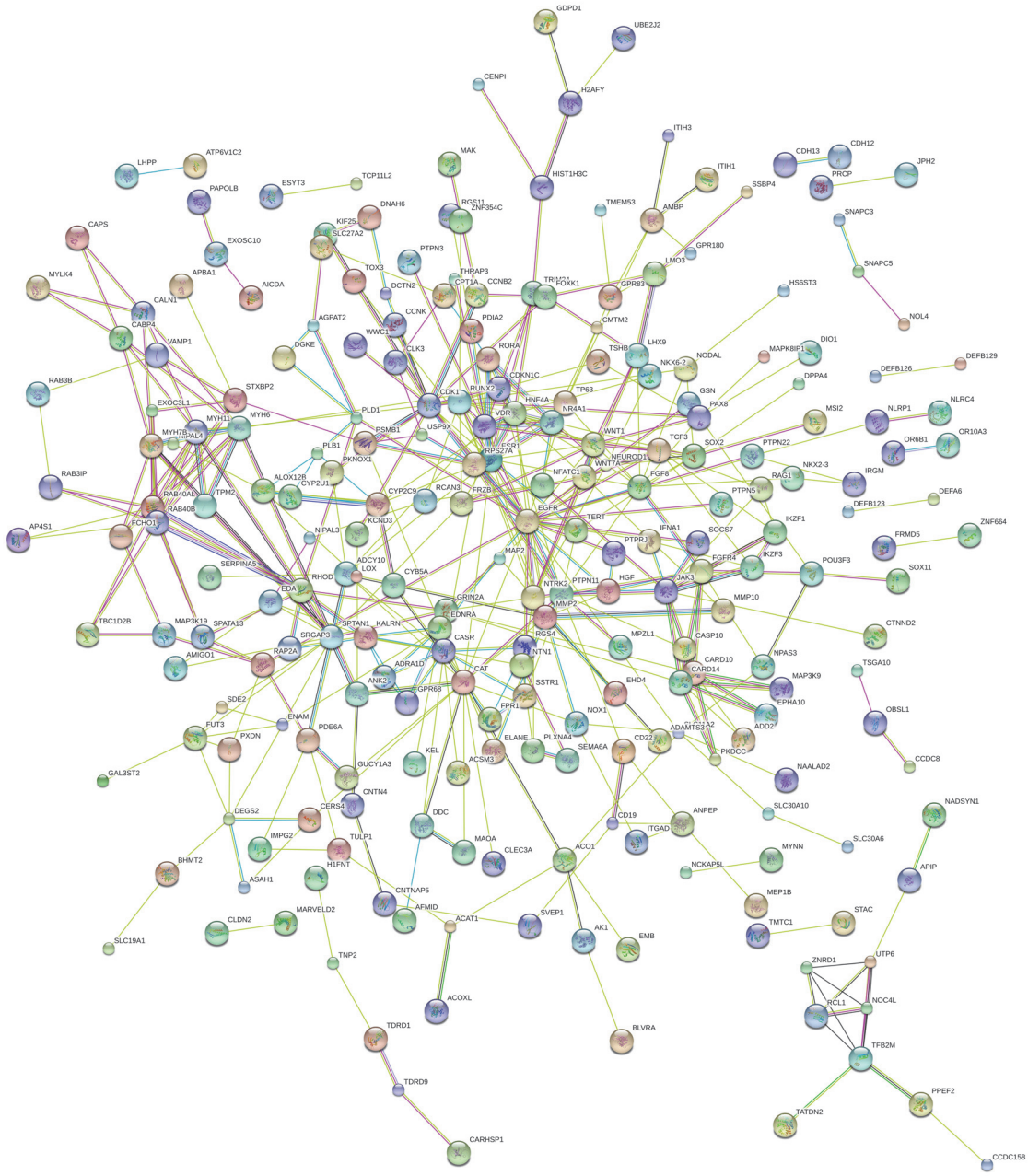
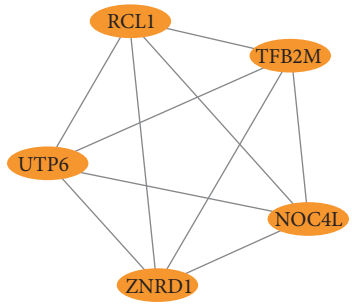


Figure 3: PPI network constructed by STRING database for the DEGs.

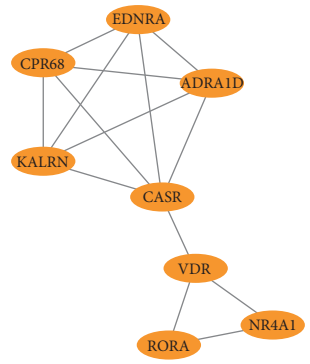
inhibition of ZNRD1/Inosine monophosphate dehydrogenase 2 (IMPDH2), upregulated DARPP-32/downregulated ZNRD1, overexpressed miR-508-5p/ZNRD1/ABCB1 activities, respectively [26–28], but further researches of ZNRD1 in osteosarcoma chemoresistance remained to be conducted. Similar to ZNRD1, MYH7B was also found upregulated in chemoresistance osteosarcoma samples. At present, MYH7B was mainly involved in pathway of cardiomyocytes, such as mitochondrial apoptosis pathway [29] but studies about cancer were rare. GPR68, a kind of pH-sensing protein, was associated with tumor cell biology, such as tumor aggressiveness by triggering the intracellular signaling cascade to promote the development of microenvironment of extracellular acidification [30, 31]. As shown in Table 1, likewise, we found the expression of GPR68 was upregulated in chemoresistance samples. Daglioglu C validated that pH-responsive Fe3O4@SiO2(FITC)-BTN/QUR/DOX multifunctional nanoparticles could potentiate the chemotherapeutic efficacy of DOX against multidrug resistance as well as counteract the survival ability of chemoresistant lung



(a)

Pathway ID	Term	Count	%	*P* value
hsa03008	Ribosome biogenesis in eukaryotes	2	40	$2.50E-02$

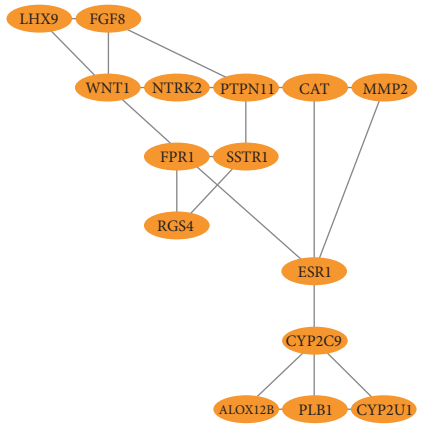
(d)



(b)

Pathway ID	Term	Count	%	*P* value
hsa04270	Vascular smooth muscle contraction	2	25	$6.71E-02$
hsa04022	cGMP-PKG signaling pathway	2	25	$9.27E-02$
hsa04020	Calcium signaling pathway	2	25	$9.97E-02$

(e)



(c)

Pathway ID	Term	Count	%	*P* value
hsa00590	Arachidonic acid metabolism	4	26.67	$1.43E-04$
hsa05205	Proteoglycans in cancer	4	26.67	$4.33E-03$
hsa00591	Linoleic acid metabolism	2	13.33	$4.93E-02$

(f)

Figure 4: The top 3 modules from the PPI network. (a) module 1, (b) module 2, (c) module 3, (d) the enriched pathways of module 1, (e) the enriched pathways of module 2, and (f) the enriched pathways of module 3.

carcinoma A549/DOX cell lines [32]. GPR68 has become an attractive target for drug development [33]. Several previous studies demonstrated that decreased CAT was highly associated with chemoresistance; for example, Xu et al. showed that intervention against miR-551b/CAT/reactive oxygen species (ROS)/Mucin-1 (MUC1) pathway might help overcome acquired chemoresistance [34]. Tumor microenvironment (TME) was characterized by hypoxia, acidosis, and dense extracellular matrix, providing tumors with resistance to various therapies, which could be effectively changed by the intravenous injection of human serum albumin (HAS)-chlorine e6 (Ce6)-CAT-paclitaxel (PTX) nanoparticles, enzyme-loaded therapeutic albumin nanoparticles. Meanwhile, H2O2 could relieve tumor hypoxia by generating oxygen within TME triggered by CAT of those nanoparticles, which made CAT a potential treatment target in various tumors [35]. Similarly, our data showed that CAT was downregulated in chemoresistance samples, which might exacerbate local microenvironment to strengthen tumor chemoresistance through the way of hypoxia, subsequent acidosis, and the like. High expression of FUT3 was proved to participate in the development of invasion, metastasis, and

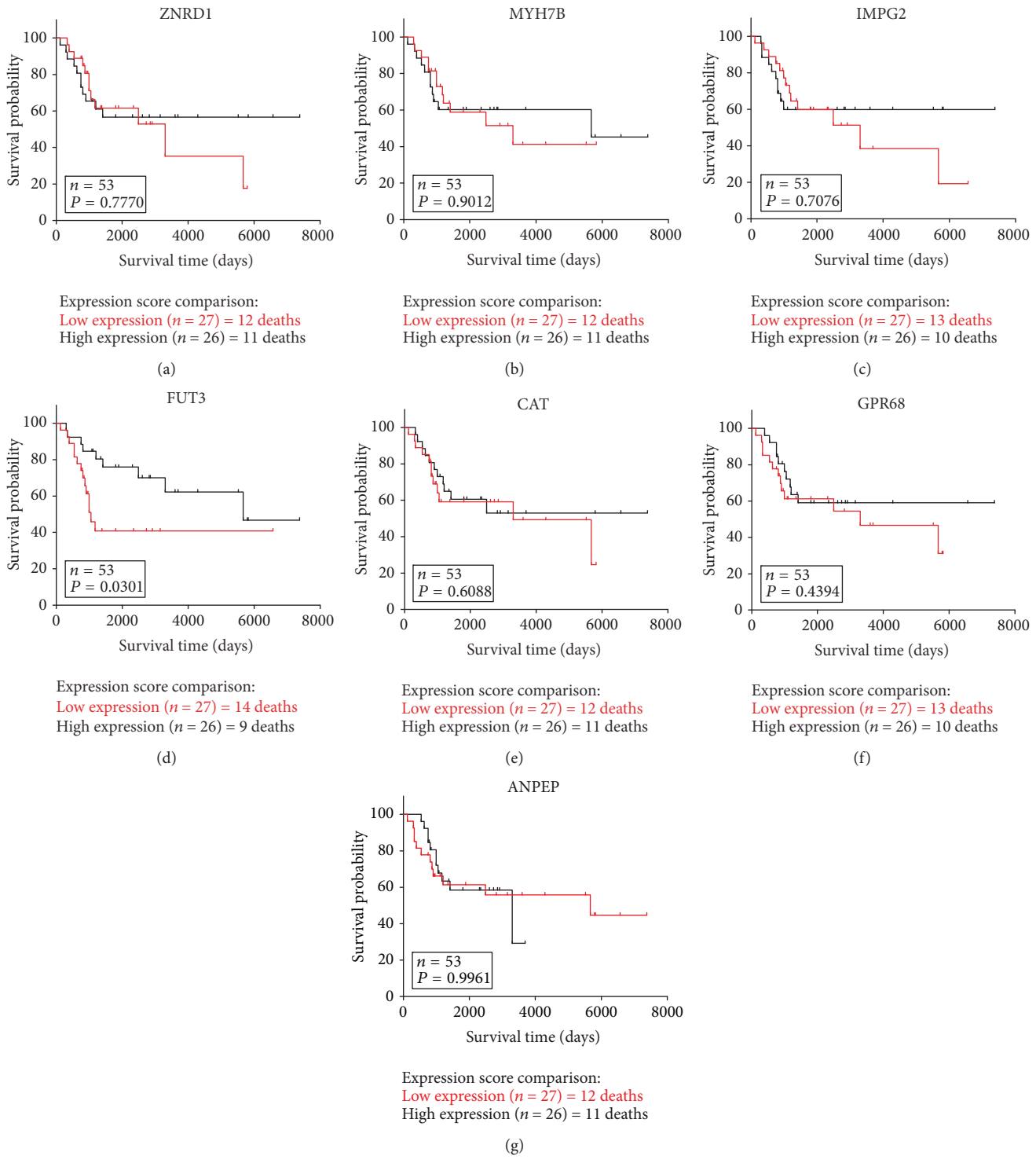


Figure 5: The survival prognostic value of hub gene in osteosarcoma from the GSE21257.

resistance to therapy by increased fucosylation activity in oral squamous cell carcinoma (OSCC), and the function could be blocked by inhibition of fucosylation [36]. In our study, FUT3 was also observed upregulated in chemoresistance samples, but it was found to be associated with a better survival prognosis (Figure 4). There were some reasons to explain that. In spite of its protumor action, death pathways were proved to be relied on fucosylation, and FUT3 was demonstrated to play an important role in natural killer-induced cytotoxicity after the recognition of sialyl Lewis X with the help of C-type lectin receptors [37, 38]. Therefore, the relationship between FUT3 and tumor was so complex

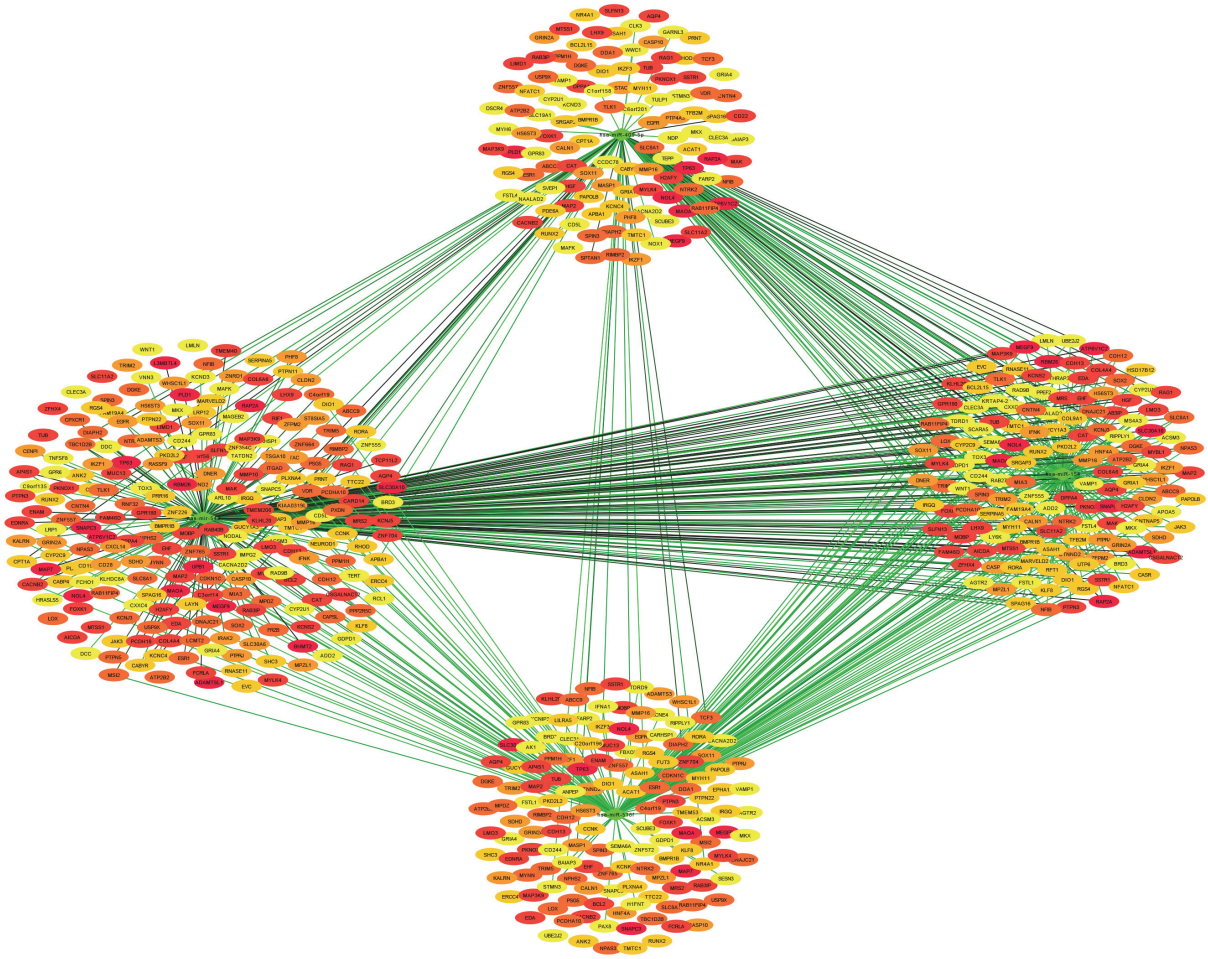


Figure 6: The network of miRNA–DEG pairs.

that little was known about its function in osteosarcoma chemoresistance. IMPG-2, a gene mainly associated with retinal disease, was upregulated in chemoresistant osteosarcoma samples in our study, but more studies involving IMPG-2 and cancer needed to be conducted [39]. Furthermore, our data showed the decreased expression of GPR180 in poor chemotherapy response samples. Homoplastically, Honda et al. found that the methylation of GPR180 was probably to encode tumor suppressors and serves as a novel prognostic marker and therapeutic target in Hepatoblastoma [40]. Based on previous studies, whether the gene GPR180 could have function in osteosarcoma by producing tumor suppressors and its concrete role in chemoresistance remained to be explored. The gene ANPEP encoded aminopeptidase N (APN). A previous study [41] showed that ANPEP was downregulated in prostate cancer (PC). On the contrary, our study showed that ANPEP expression of good chemotherapy response samples was approximately two times that in chemoresistance osteosarcoma samples. However, the difference caused by the types of tumors or chemoresistance needed to be further studied. In urothelial carcinoma, APN could increase cytotoxicity of melphalan-flufenamide to play anticancer effect by amplifying the intracellular loading of melphalan [42]. The studies of chemoresistance associated with ANPEP have not been conducted so far, but Viktorsson et al. [42] offered researchers a new way for treatment, which made ANPEP a significant therapeutic target. Several researches in chemoresistance-associated fields have already demonstrated that CDK1 participated in the development of chemoresistance in pathways, such as GRP78/CDC2 [43]. Likewise, in our study, the expression of CDK1 was increased. Hayashi et al. found increased DNA repair activity in the G2-M transition promoted temozolomide (TMZ) resistance and CDC2 inhibitor flavopiridol (FP) treatment could resensitize TMZ-resistant clones in a p53-independent manner in glioma cells [44]. Besides, the combination of ERK inhibitor PD98059 and Taxol could improve the sensitivity of taxol-resistant tumor cells with the decreased CDC2 activity [45].

Compared with that of good response, 5 DEMs were acquired in GSE30934 in poor chemotherapy response samples (Table 2). Among them, hsa-miR-543 and hsa-miR-518f

were found to have a relation to ZNRD1 and CAT, respectively. In our results, hsa-miR-543 was downregulated in chemoresistant samples. Previous study in this field was limited. In other aspects, decreased expression of it was involved in osteosarcoma angiogenesis which might be caused by connective tissue growth factor (CTGF) in phospholipase C (PLC)/protein kinase C (PKCδ) signaling pathway. Besides, hsa-miR-543 was also proved to be linked with tumor staging [46], cell proliferation, and the glycolytic pathway [47]. The studies between chemoresistance-promoted gene ZNRD1 and hsa-miR-543 have not been conducted yet, but biological functions mentioned above made hsa-miR-543 become an important therapeutic target. Moreover, the role of hsa-miR-518f in chemoresistance or the development of tumor was rarely known. Hsa-miR-518e and hsa-miR-518b, homologues of hsa-miR-518f, were demonstrated to be upregulated in hepatocellular carcinoma (HCC) [48]. Besides, a previous study showed that has-miR-518c-5p could regulate the growth and metastasis of oral cancer [49]. Consequently, further research to hsa-miR-518f was of great importance.

In summary, we identified 668 DEGs and 5 DEMs from GEO2R between good chemotherapy response samples and poor chemoresistance samples in osteosarcoma. And many of them, such as ZNRD1, GPR68, CAT, FUT3, ANPEP, CDK1, and hsa-miR-543, might be key genes related to osteosarcoma chemoresistance. These findings provided a series of promising treatment targets and enlightened us on the further investigations of the molecular mechanisms.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

References

[1] L. Mirabello, R. J. Troisi, and S. A. Savage, "International osteosarcoma incidence patterns in children and adolescents, middle ages and elderly persons," *International Journal of Cancer*, vol. 125, no. 1, pp. 229–234, 2009.

[2] L. Mirabello, R. J. Troisi, and S. A. Savage, "Osteosarcoma incidence and survival rates from 1973 to 2004: data from the surveillance, epidemiology, and end results program," *Cancer*, vol. 115, no. 7, pp. 1531–1543, 2009.

[3] H. D. Dorfman and B. Czerniak, "Bone cancers," *Cancer*, vol. 75, no. 1, supplement, pp. 203–210, 1995.

[4] K. P. Anfinsen, S. S. Devesa, F. Bray et al., "Age-period-cohort analysis of primary bone cancer incidence rates in the United States (1976-2005)," *Cancer Epidemiology, Biomarkers & Prevention*, vol. 20, no. 8, pp. 1770–1777, 2011.

[5] P. C. Valery, M. Laversanne, and F. Bray, "Bone cancer incidence by morphological subtype: a global assessment," *Cancer Causes & Control*, vol. 26, no. 8, pp. 1127–1139, 2015.

[6] L. Mirabello, R. Pfeiffer, G. Murphy et al., "Height at diagnosis and birth-weight as risk factors for osteosarcoma," *Cancer Causes & Control*, vol. 22, no. 6, pp. 899–908, 2011.

[7] J. R. B. Musselman, T. L. Bergemann, J. A. Ross et al., "Case-parent analysis of variation in pubertal hormone genes and pediatric osteosarcoma: a children's oncology group (COG) study," *International Journal of Molecular Epidemiology and Genetics*, vol. 3, no. 4, pp. 286–293, 2012.

[8] M. Kansara, M. W. Teng, M. J. Smyth, and D. M. Thomas, "Translational biology of osteosarcoma," *Nature Reviews Cancer*, vol. 14, no. 11, pp. 722–735, 2014.

[9] L. Mirabello, K. Yu, S. I. Berndt et al., "A comprehensive candidate gene approach identifies genetic variation associated with osteosarcoma," *BMC Cancer*, vol. 11, Article ID 209, 2011.

[10] S. H. Orkin, D. E. Fisher, A. T. Look, S. Lux, D. Ginsburg, and D. G. Nathan, *Oncology of Infancy and Childhood*, Elsevier Health Sciences, 2009.

[11] D. C. Dahlin and M. B. Coventry, "Osteogenic sarcoma, a study of six hundred cases," *The Journal of Bone & Joint Surgery*, vol. 49, no. 1, pp. 101–110, 1967.

[12] R. C. Marcove, V. Miké, J. V. Hajek, A. G. Levin, and R. V. Hutter, "Osteogenic sarcoma under the age of twenty-one," *The Journal of Bone & Joint Surgery*, vol. 52, no. 3, pp. 411–423, 1970.

[13] P. A. Meyers, G. Heller, J. Healey et al., "Chemotherapy for non-metastatic osteogenic sarcoma: the memorial sloan-kettering experience," *Journal of Clinical Oncology*, vol. 10, no. 1, pp. 5–15, 1992.

[14] A. M. Goorin, D. J. Schwartzentruber, M. Devidas et al., "Presurgical chemotherapy compared with immediate surgery and adjuvant chemotherapy for nonmetastatic osteosarcoma: Pediatric Oncology Group Study POG-8651," *Journal of Clinical Oncology*, vol. 21, no. 8, pp. 1574–1580, 2003.

[15] B. Kempf-Bielack, S. S. Bielack, H. Jürgens et al., "Osteosarcoma relapse after combined modality therapy: an analysis of unselected patients in the Cooperative Osteosarcoma Study Group (COSS)," *Journal of Clinical Oncology*, vol. 23, no. 3, pp. 559–568, 2005.

[16] H. I. Vos, M. J. H. Coenen, H.-J. Guchelaar, and D. M. W. M. te Loo, "The role of pharmacogenetics in the treatment of osteosarcoma," *Drug Discovery Therapy*, vol. 21, no. 11, pp. 1775–1786, 2016.

[17] M. Serra and C. M. Hattinger, "The pharmacogenomics of osteosarcoma," *The Pharmacogenomics Journal*, vol. 17, no. 1, pp. 11–20, 2017.

[18] V. N. Kim, J. Han, and M. C. Siomi, "Biogenesis of small RNAs in animals," *Nature Reviews Molecular Cell Biology*, vol. 10, no. 2, pp. 126–139, 2009.

[19] D. P. Bartel, "MicroRNAs: target recognition and regulatory functions," *Cell*, vol. 136, no. 2, pp. 215–233, 2009.

[20] M. A. Valencia-Sanchez, J. Liu, G. J. Hannon, and R. Parker, "Control of translation and mRNA degradation by miRNAs and siRNAs," *Genes & Development*, vol. 20, no. 5, pp. 515–524, 2006.

[21] P. Gaudet, N. Škunca, J. C. Hu, and C. Dessimoz, "Primer on the Gene Ontology," in *The Gene Ontology Handbook*, vol. 1446 of *Methods in Molecular Biology*, pp. 25–37, Springer New York, New York, NY, USA, 2017.

[22] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, "KEGG: new perspectives on genomes, pathways, diseases and drugs," *Nucleic Acids Research*, vol. 45, no. 1, pp. D353–D361, 2017.

[23] D. Szklarczyk, A. Franceschini, S. Wyder et al., "STRING v10: protein-protein interaction networks, integrated over the tree of life," *Nucleic Acids Research*, vol. 43, pp. D447–D452, 2015.

[24] P. Shannon, A. Markiel, O. Ozier et al., "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Research*, vol. 13, no. 11, pp. 2498–2504, 2003.

[25] G. A. Calin, C. Sevignani, C. D. Dumitru et al., "Human microRNA genes are frequently located at fragile sites and

genomic regions involved in cancers," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 101, no. 9, pp. 2999–3004, 2004.

[26] L. Hong, T. Qiao, Y. Han et al., "ZNRD1 mediates resistance of gastric cancer cells to methotrexate by regulation of IMPDH2 and Bcl-2," *The International Journal of Biochemistry & Cell Biology*, vol. 84, no. 2, pp. 199–206, 2006.

[27] L. Hong, Y. Zhao, J. Wang et al., "Reversal of multidrug resistance of adriamycin-resistant gastric adenocarcinoma cells through the up-regulation of DARPP-32," *Digestive Diseases and Sciences*, vol. 53, no. 1, pp. 101–107, 2008.

[28] Y. Shang, B. Feng, L. Zhou et al., "The miR27b-CCNG1-P53-miR-508-5p axis regulates multidrug resistance of gastric cancer," *Oncotarget*, vol. 7, no. 1, pp. 538–549, 2016.

[29] J. Wang, Z. Jia, C. Zhang et al., "miR-499 protects cardiomyocytes from H_2O_2-induced apoptosis via its effects on *Pdcd4* and *Pacs2*," *RNA Biology*, vol. 11, no. 4, pp. 339–350, 2014.

[30] W.-C. Huang, P. Swietach, R. D. Vaughan-Jones, O. Ansorge, and M. D. Glitsch, "Extracellular acidification elicits spatially and temporally distinct Ca2+ signals," *Current Biology*, vol. 18, no. 10, pp. 781–785, 2008.

[31] H. Saxena, D. A. Deshpande, B. C. Tiegs et al., "The GPCR OGR1 (GPR68) mediates diverse signalling and contraction of airway smooth muscle in response to small reductions in extracellular pH," *British Journal of Pharmacology*, vol. 166, no. 3, pp. 981–990, 2012.

[32] C. Daglioglu, "Enhancing tumor cell response to multidrug resistance with pH-sensitive quercetin and doxorubicin conjugated multifunctional nanoparticles," *Colloids and Surfaces B: Biointerfaces*, vol. 156, pp. 175–185, 2017.

[33] X.-P. Huang, J. Karpiak, W. K. Kroeze et al., "Allosteric ligands for the pharmacologically dark receptors GPR68 and GPR65," *Nature*, vol. 527, no. 7579, pp. 477–483, 2015.

[34] X. Xu, A. Wells, M. T. Padilla, K. Kato, K. C. H. Kim, and Y. Lin, "A signaling pathway consisting of miR-551b, catalase and MUC1 contributes to acquired apoptosis resistance and chemoresistance," *Carcinogenesis*, vol. 35, no. 11, pp. 2457–2466, 2014.

[35] Q. Chen, J. Chen, C. Liang et al., "Drug-induced co-assembly of albumin/catalase as smart nano-theranostics for deep intratumoral penetration, hypoxia relieve, and synergistic combination therapy," *Journal of Controlled Release*, vol. 263, pp. 79–89, 2017.

[36] V. Desiderio, P. Papagerakis, V. Tirino et al., "Increased fucosylation has a pivotal role in invasive and metastatic properties of head and neck cancer stem cells," *Oncotarget*, vol. 6, no. 1, pp. 71–84, 2015.

[37] K. Higai, A. Ichikawa, and K. Matsumoto, "Binding of sialyl Lewis X antigen to lectin-like receptors on NK cells induces cytotoxicity and tyrosine phosphorylation of a 17-kDa protein," *Biochimica et Biophysica Acta (BBA)—General Subjects*, vol. 1760, no. 9, pp. 1355–1363, 2006.

[38] C. Ohyama, S. Kanto, K. Kato et al., "Natural killer cells attack tumor cells expressing high levels of sialyl Lewis x oligosaccharides," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 99, no. 21, pp. 13789–13794, 2002.

[39] D. Bandah-Rozenfeld, R. W. J. Collin, E. Banin et al., "Mutations in IMPG2, Encoding interphotoreceptor matrix proteoglycan 2, cause autosomal-recessive retinitis pigmentosa," *American Journal of Human Genetics*, vol. 87, no. 2, pp. 199–208, 2010.

[40] S. Honda, M. Minato, H. Suzuki et al., "Clinical prognostic value of DNA methylation in hepatoblastoma: four novel tumor suppressor candidates," *Cancer Science*, vol. 107, no. 6, pp. 812–819, 2016.

[41] K. D. Sørensen, M. O. Abildgaard, C. Haldrup et al., "Prognostic significance of aberrantly silenced ANPEP expression in prostate cancer," *British Journal of Cancer*, vol. 108, no. 2, pp. 420–428, 2013.

[42] K. Viktorsson, C.-H. Shah, T. Juntti et al., "Melphalan-flufenamide is cytotoxic and potentiates treatment with chemotherapy and the Src inhibitor dasatinib in urothelial carcinoma," *Molecular Oncology*, vol. 10, no. 5, pp. 719–734, 2016.

[43] W. Li, W. Wang, H. Dong et al., "Cisplatin-induced senescence in ovarian cancer cells is mediated by GRP78," *Oncology Reports*, vol. 31, no. 6, pp. 2525–2534, 2014.

[44] T. Hayashi, K. Adachi, S. Ohba, and Y. Hirose, "The Cdk inhibitor flavopiridol enhances temozolomide-induced cytotoxicity in human glioma cells," *Journal of Neuro-Oncology*, vol. 115, no. 2, pp. 169–178, 2013.

[45] X. Lin, Y. Liao, X. Chen, D. Long, T. Yu, and F. Shen, "Regulation of oncoprotein 18/Stathmin signaling by ERK concerns the resistance to Taxol in nonsmall cell lung cancer cells," *Cancer Biotherapy and Radiopharmaceuticals*, vol. 31, no. 2, pp. 37–43, 2016.

[46] L.-H. Wang, H.-C. Tsai, Y.-C. Cheng et al., "CTGF promotes osteosarcoma angiogenesis by regulating miR-543/angiopoietin 2 signaling," *Cancer Letters*, vol. 391, pp. 28–37, 2017.

[47] H. Zhang, X. Feng, T. Wang et al., "MiRNA-543 promotes osteosarcoma cell proliferation and glycolysis by partially suppressing PRMT9 and stabilizing HIF-1α protein," *Oncotarget*, vol. 8, no. 2, pp. 2342–2355, 2017.

[48] W. Wang, L.-J. Zhao, Y.-X. Tan, H. Ren, and Z.-X. Qi, "MiR-138 induces cell cycle arrest by targeting cyclin D3 in hepatocellular carcinoma," *Carcinogenesis*, vol. 33, no. 5, pp. 1113–1120, 2012.

[49] M. Kinouchi, D. Uchida, N. Kuribayashi et al., "Involvement of miR-518c-5p to growth and metastasis in oral cancer," *PLoS ONE*, vol. 9, no. 12, Article ID e115936, 2014.

Hindawi
BioMed Research International
Volume 2018, Article ID 3453706, 7 pages
https://doi.org/10.1155/2018/3453706

Research Article

High Mobility Group Box Protein 1 Serves as a Potential Prognostic Marker of Lung Cancer and Promotes Its Invasion and Metastasis by Matrix Metalloproteinase-2 in a Nuclear Factor-κB-Dependent Manner

Xiaojin Wu,[1] Weitao Wang,[2] Yuanyuan Chen,[1] Xiangqun Liu,[3] Jindong Wang,[4] Xiaobin Qin,[5] Dawei Yuan,[2] Tao Yu,[5] Guangxia Chen,[6] Yanyan Mi,[7] Jie Mou,[7] Jinpeng Cui,[8] Ankang Hu,[9] Yunxiang E,[5] and Dongsheng Pei[10]

[1] *Department of Radiation Oncology, The First People's Hospital of Xuzhou, Xuzhou, Jiangsu 221002, China*
[2] *Geneis Beijing Co., Ltd., Beijing 100102, China*
[3] *Department of Respiratory Diseases, The First People's Hospital of Xuzhou, Xuzhou, Jiangsu 221002, China*
[4] *Department of Chest Surgery, The First People's Hospital of Xuzhou, Xuzhou, Jiangsu 221002, China*
[5] *Department of Tumor, The First People's Hospital of Xuzhou, Xuzhou, Jiangsu 221002, China*
[6] *Department of Gastroenterology, The First People's Hospital of Xuzhou, Xuzhou, Jiangsu 221002, China*
[7] *Department of Pharmacy, Xuzhou Medical College, Xuzhou, Jiangsu 221004, China*
[8] *Clinical Laboratory of Yantaishan Hospital, No. 91, Jiefang Road, Yantai, Shandong 264001, China*
[9] *Laboratory Animal Center, Xuzhou Medical College, Xuzhou, Jiangsu 221004, China*
[10] *Department of Pathology, Xuzhou Medical College, Xuzhou, Jiangsu 221004, China*

Correspondence should be addressed to Dongsheng Pei; dspei@xzhmu.edu.cn

Received 9 October 2017; Revised 1 December 2017; Accepted 4 February 2018; Published 19 April 2018

Academic Editor: Tao Huang

Several studies have reported a significant role of high mobility group box protein 1 (HMGB1) in lung cancer. Nevertheless, there is a lack of knowledge regarding the expression of HMGB1 and its correlation with the clinicopathological features of lung cancer. In addition, the potential molecular mechanisms underlying the role of HMGB1 in lung cancer are still unknown. We therefore investigated the clinicopathological and prognostic significance as well as the potential role of HMGB1 in the development and progression of lung cancer. HMGB1 expression in the tumor tissues of the cohort correlated with clinicopathological features. Moreover, lung cell migration and invasion were significantly increased after treatment with HMGB1. The matrix metalloproteinase-2 (MMP-2) expression and activity were upregulated after treatment with HMGB1, while the upregulated expression of MMP-2 stimulated by HMGB1 in lung cancer cells was significantly reduced with the blockage of si-p65. These results indicated that HMGB1 expression was significantly associated with lung cancer progression. We also showed that HMGB1 promoted lung cancer invasion and metastasis by upregulating the expression and activity of MMP-2 in an NF-κB-dependent manner. Taken together, these data suggested that HMGB1 may be a potential prognosis and therapeutic marker for lung cancer.

1. Introduction

Lung cancer has currently become one of the most serious diseases threatening human health. Its incidence and mortality rates are the highest of all malignant tumors in the world [1–4]. Non-small cell lung cancer (NSCLC) accounts for 80%–85% of all lung cancers that involve a multifactorial, multistep, and multistage complex process [5–9]. Although significant progress has been made in clinical treatments with the rapid development of molecular biology techniques and clinical treatments, the survival of lung cancer patients is still limited. It is therefore important to investigate the molecular

mechanisms of invasion and metastasis of lung cancer and to identify potential prognostic markers.

High mobility group box protein 1 (HMGB1) is a non-histone, chromatin-binding nuclear protein that is widely found in eukaryotic cells, with a protein structure containing two homologous DNA-binding domains (A and B boxes) and a C-terminal tail that participates in transcription, DNA preparation, cell growth and differentiation, and extracellular signal transduction [10–12]. HMGB1 has been reported to be involved in cancer development, invasion, and metastases. In addition, HMGB1 is frequently highly expressed in various malignant tumors and is an early marker for these tumors [13–17]. In recent years, various studies to characterize the expression of HMGB1 in NSCLC patients have suggested that HMGB1 plays a vital role in the diagnosis and prognosis of NSCLC [12, 18–24]. However, this suggestion is still controversial. Shen et al. [22] suggested that HMGB1 is downregulated in NSCLC tissues, while other studies reported that the level of HMGB1 is upregulated in NSCLC tissues compared with that of normal tissues [12, 20, 24]. Thus, further research is needed to clarify the possible clinical diagnostic and prognostic values of HMGB1 in NSCLC.

It is well-known that invasion and metastasis, which are the main causes of high mortality, are common characteristics of malignant tumors. Extracellular HMGB1 is thought to induce cancer cell growth, mobility, invasion, and metastasis via binding to specific membrane receptors, including the receptor for advanced glycation end products (RAGE), and blockage of the RAGE-HMGB1 complex suppresses tumor growth and metastasis [20, 25–27]. Taguchi et al. also reported that the metastasis of Lewis lung tumor cells was inhibited when treated with anti-HMGB1 antibody, suggesting that HMGB1 is associated with tumor invasion and metastasis [26].

Matrix metalloproteinase 2 (MMP2) is a member of the Ca^{2+}- and Zn^{2+}-dependent endogenous protease family (MMPs), whose expression in tumors is reported to be correlated with carcinoma invasion and metastasis [28–30]. MMP2 expression is upregulated in many cancers such as in glioblastomas, melanomas, breast cancers, and colon cancers [30]. It has also been reported that the high MMP2 expression in NSCLC is an independent prognostic factor that is closely related to its clinical stage, pathological grade, lymphatic metastasis, and prognosis [19].

As a key transcription factor, the nuclear factor (NF)-κB plays an important role in tumor development and progression [31]. Activated NF-κB can trigger chemokine and cytokine production, which further results in tissue damage [32]. Moreover, activated NF-κB signaling pathways have been correlated with MMP2 expression, and there is evidence that deacetylation of NF-κB reduces MMP2 expression, leading to the inhibition of NSCLC cell invasion [33].

However, the underlying mechanisms involving HMGB1 promotion of lung cancer invasion and metastasis still remain unclear and require further elucidation. In the present study, our aim was to characterize the association between the expression of HMGB1 and the clinicopathological and prognostic factors of NSCLC and to investigate how HMGB1 promotes lung cancer invasion and metastasis.

2. Materials and Methods

2.1. Materials. The lung cancer tissue chip was purchased from Shanghai core Biological Technology Co., Ltd., number HLug-Ade180Sur-01. In 90 cases of lung cancer tissues and 90 cases of lung adjacent noncancerous tissues, TNM stage was I–III stages; lymph node metastasis was N0 in 39 cases, N1–N3 in 36 cases, and N × 15 cases in other cases. The clinical data and follow-up data were complete. The operation time was 2004.7–2009.6; the follow-up time was 2014.8, followed up for 5–10 years.

A549 and H1299 cell lines were purchased from the US standard library (American Type Culture Collection, ATCC).

2.2. Immunohistochemical Staining. The immunohistochemical staining method we used was Envision™, performed as described previously [34]. As for tissue immunohistochemistry staining results analysis, lung cancer tissue staining results were evaluated by IRS score under double-blind conditions by two senior pathological experts. The average staining score was calculated by combining the positive staining intensity and the percentage of positive cells. Dyeing strength was as follows: nonstaining, 0 points; weak coloring, 1 points; medium coloring, 2 points; strong coloring, 3 points. The percentage of positive cells ranged from 0 to 25%, 1 points; 26% to 50%, 2 points; 51% to 75%, 3 points; 76% to 100%, 4 points. Two points multiplied by the following scores: 0 for negative; 1–3 for weak positive; 4–6 for medium positive; and 8–12 for strong positive. We stipulated 4 points and below for HMGB1 low expression and 4 points above for HMGB1 high expression.

2.3. In Vitro Invasion Assay. The in vitro invasion assays were carried out using Transwell® chamber with 6.5 mm diameter polycarbonate filters as described previously [35]. HMGB1 pretreated lung cancer cells digested by trypsin were suspended by serum-free medium and then counted. The concentrations of A549 and H1299 were 4.5×10^4/ml and 3×10^4/ml, respectively. 10% serum was added to the bottom of the 24-orifice plate and cell suspension was added to the incubation chamber for 24 hours. We made the methanol fixation at RT for 30 min and then the PBS wash twice. Crystal violet was dyed for 15~30 minutes and 5 pictures were taken at random under the microscope and counted.

2.4. Gelatin Zymography. Gelatin zymography was performed as described previously [36]. We collected the supernatant of HMGB1 pretreated or not lung cancer cells, measured the proteins concentration, and then mixed it with 4x loading buffer about 30–50 ug/40 ul. SDS-PAGE electrophoresis was performed at 100 V for about 1.5 hours. Low speed oscillation elution was carried out in 2.5% Triton X-100 eluted with SDS (elution for 90 minutes), followed by rinsing with distilled water and then gel incubation at 37° for 60 h until a transparent strip appears. This was then stained at RT for 1–4 h and then destained for 2–5 h to observe the MMP2 strip.

Table 1: HMGB1 expression level in lung cancer tissues and normal tissues adjacent to lung cancers.

Variables	HMGB1 staining			P^*
	Low (%)	High (%)	Total	
Lung cancer tissues	39 (43.3)	51 (56.7)	90	0.001
Adjacent noncancerous tissues	63 (70.0)	27 (30.0)	90	

*The Student t-test for P value.

2.5. Western Blot Analysis. Protein level was detected by western blot analysis as described previously [37]. After treatment with HMGB-1, A549 and H1299 cells were washed twice by ice-cold PBS. Cells were lysed by RIPA lysis buffer and then proteins were extracted. Denaturation of equal amount of supernatant protein in boiling SDS sample buffer was performed and then the samples were subjected to 10% SDS-PAGE. After that, polyvinylidene difluoride membranes were used for the proteins transfer. 5% dry skim milk was used for membrane blocking and then membranes were incubated with primary antibodies p65, MMP2. In this experiment, β-actin was used as an internal reference for the reliability of the experiment. Finally, membranes were treated with enhanced chemiluminescent system for visualization of the protein bands. The bands were quantified using Image J software.

2.6. Statistical Analysis. The correlation analysis of the expression level of HMGB1 and various clinicopathologic parameters was made using Fisher exact test. The survival data were analyzed using Kaplan–Meier survival curve analysis and Log-rank test. Measurement data represents mean ± standard deviation ($x \pm s$). The means of two groups were compared with t test. We used SPSS18.0 statistical software for data processing. $P < 0.05$ shows that the difference was statistically significant, $P < 0.01$ shows that there is a significant difference.

3. Results

3.1. Expression of HMGB1 in Lung Cancer Tissues and Normal Tissues Adjacent to the Lung Cancer. Immunohistochemical staining showed that the HMGB1 high expression percentages in lung cancer tissues and normal tissues adjacent to the lung cancer tissues from 90 patients were 56.7% and 30.0%, respectively, which showed that the level of HMGB1 expression in lung cancer tissue was significantly higher than in the tissue adjacent to the lung cancers (Table 1).

3.2. The Relationship between the Expression of HMGB1 in Lung Cancer Tissue and Clinical Pathological Parameters. HMGB1 expression level was associated with the clinical pathological characteristics of the patients with lung cancer, such as T stage ($P = 0.027$) and lymph node metastasis of the tumor ($P = 0.019$), and was closely related to the clinical stage ($P = 0.012$). However, when compared with the patients' gender, age, and size, there was no significant difference ($P > 0.05$) (Table 2).

3.3. The Relationship between the Level of HMGB1 Expression of Lung Cancer Tissue and the Prognoses of Patients. In order to investigate whether the high expression of HMGB1 was correlated with the poor prognosis of patients with lung cancers, we analyzed the 5-year survival rates of 90 cases of lung cancer patients using Kaplan–Meier survival curves. According to the staining intensity, we chose different time nodes using the survival analyses, which showed that the postoperative total survival percentage of high HMGB1 expression was significantly lower than low HMGB1 expression for lung cancer patients ($P = 0.0011$, HR = 0.0011, 95% CI: 0.2623 0.7166; Figure 1). Overall, the results showed that the high expression of HMGB1 was closely related to the poor prognoses of patients with lung cancer.

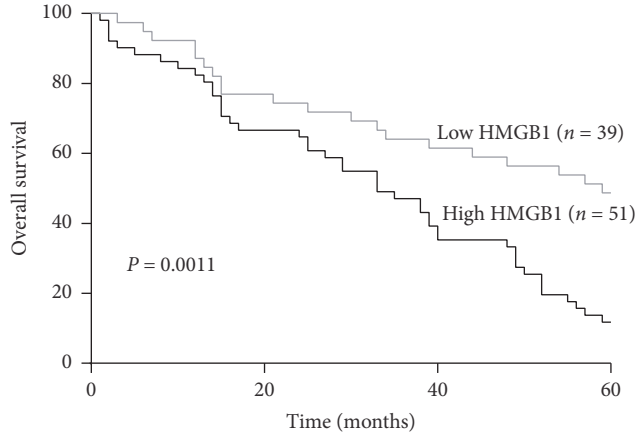


Figure 1: High mobility group box protein 1 expression correlated with a poorer 5-year survival for 90 lung cancer patients *($P = 0.0011$)*. Significant differences were assessed by the Student t-test.

3.4. HMGB1 Promotes Lung Cancer Cell Invasion. In order to determine whether HMGB1 promoted lung cancer cell invasion, the lung cancer cell lines, A549 and H1299, were pretreated with HMGB1, and the Transwell assay was used to determine the changes of cell invasion. Compared with the control group, the number of A549 and H1299 cells passing through the Transwell chambers increased by 46% and 41%, respectively, after pretreatment with HMGB1. The increased invasiveness was significantly different, suggesting that treatment with HMGB1 increases the invasive ability of lung cancer cells (Figure 2).

3.5. HMGB1 Activates MMP2 Enzyme Activity. To further study the mechanism of HMGB1-promoting lung cancer cell invasion, gelatin zymography assays were performed in lung cancer cell lines, A549 and H1299, after pretreatment with HMGB1. The results showed that MMP2 enzyme activities in the A549 and H1299 cells were significantly higher after treatment with HMGB1 than the control group (Figure 3).

3.6. HMGB1 Promotes MMP-2 Expression via the NF-κB Pathway. Our previous study showed that HMGB1 induced NF-κB expression in human bronchial epithelial cells in a dose-dependent manner. To investigate the mechanism of HMGB1 promotion of MMP2 expression, western blots were used to detect NF-κB expression. The results showed

Table 2: The relationship between high mobility group box protein 1 staining and the clinicopathological characteristics of 90 lung cancer patients.

Variables	HMGB-1 staining			
	Low (%)	High (%)	Total	P^*
Age				
≥60 years	23 (41.8)	32 (58.2)	55	0.828
<60 years	16 (45.7)	19 (54.3)	35	
Gender				
Male	19 (38.8)	30 (61.2)	49	0.396
Female	20 (48.8)	21 (51.2)	41	
Tumor size				
≥4 cm	23 (44.2)	28 (55.8)	51	0.832
<4 cm	16 (41.0)	23 (59.0)	39	
pT status				
pT_1-pT_2	34 (50.7)	33 (49.3)	67	0.027
pT_3-pT_4	5 (21.7)	18 (78.3)	23	
pN status				
pN_0	22(56.4)	17(43.6)	39	0.019
pN_1–pN_3	10(27.8)	26(72.2)	36	
TNM stage				
I-II	35 (50.7)	34 (49.3)	69	0.012
III	4 (19.0)	17 (81.0)	21	

*The Student t-test for P value.

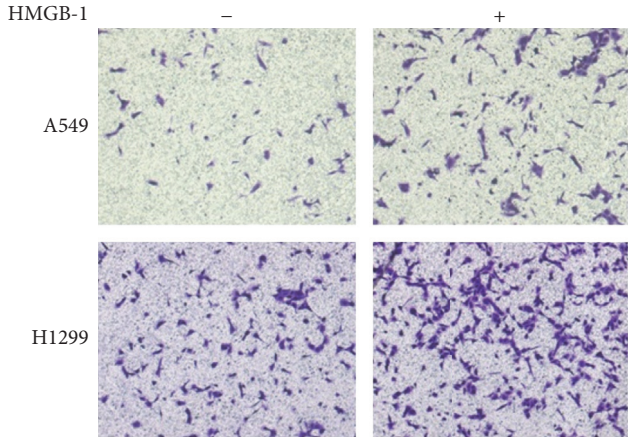


Figure 2: High mobility group box protein 1 (HMCB1) promoted invasion of lung cancer cells. Pretreatment of HMGB1 significantly promoted invasion of A549 and H1299 cells.

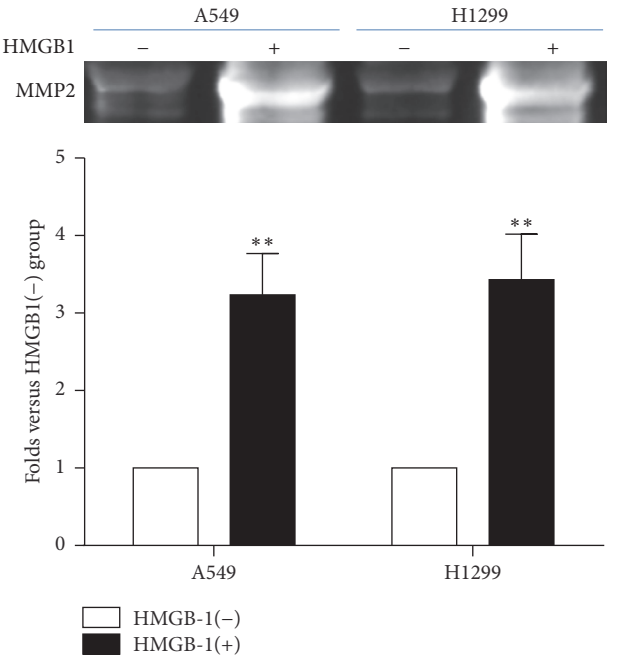


Figure 3: HMGB1 promoted the activity of MMP2 in lung cancer cells. Pretreatment of HMGB1 significantly promoted the activity of MMP2 in A549 and H1299 cells (**$P < 0.01$ compared with the control). Significant differences were assessed by the Student t-test.

significantly increased NF-κB expression of the HMGB1 pretreatment group versus the control group (Figure 4). In addition, p65 siRNA was transfected into the A549 and H1299 lung cancer cell lines, together with HMGB1 treatment. Compared with the control group, treatment with HMGB1 promoted MMP2 expression when there was no p65 siRNA transfection, while the expression of MMP2 was significantly decreased in the p65 siRNA transfection groups compared with the p65 siRNA control group (Figure 5). These results showed that HMGB1 increased the expression of MMP2, thus promoting the invasive ability of lung cancer cells via the NF-κB pathway.

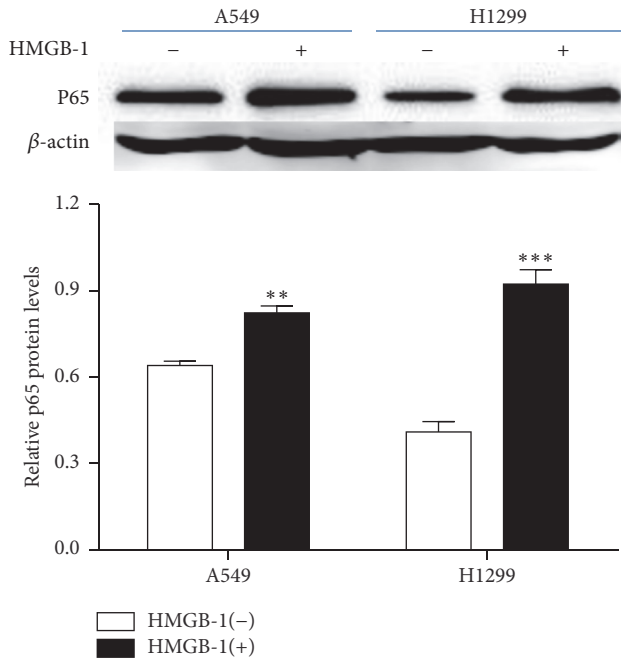


Figure 4: Western blot analysis of the expression level of p65 in lung cancer cells after treatment with HMGB1. Pretreatment of HMGB1 significantly increased the expression level of p65 in A549 and H1299 cells. ($^{**}P < 0.01$, $^{***}P < 0.001$ compared with the control). Significant differences were assessed by the Student t-test.

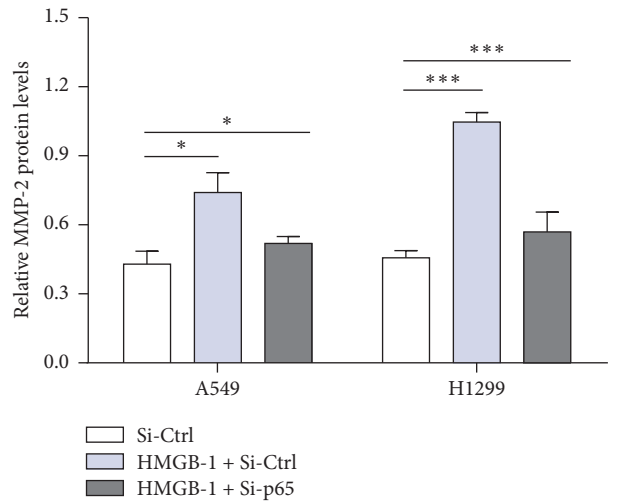


Figure 5: Western blot analysis of the expression level of MMP2 in lung cancer cells after SiRNA transfection. Pretreatment of HMGB1 significantly increased the expression level of MMP2 in A549 and H1299 cells. However, inhibition of p65 significantly attenuated the expression level of MMP2 in A549 and H1299 cells. ($^{*}P < 0.05$, $^{***}P < 0.001$ compared with the control). Significant differences were assessed by the Student t-test.

4. Discussion

Little is known about the pathogenesis of NSCLC and the identification of effective markers for early diagnosis and prognosis of NSCLC. Therefore, identification of effective markers for early diagnosis and prognosis of NSCLC has become an important objective of recent studies. HMGB1 has been reported to be associated with the progression of NSCLC. Recent studies have shown that HMGB1 can participate in cell differentiation, migration, regeneration, and mediation of inflammation, particularly in binding to RAGE, to affect tumor growth, invasion, and metastasis [38]. Shang et al. first reported that serum HMGB1 levels were significantly increased in patients with lung cancer when compared with control subjects [20]. In the present study, we used tumor tissue microarrays to characterize an independent cohort of 90 NSCLC patients for the expression of HMGB1. The clinical analyses showed a significantly higher expression of HMGB1 in the lung cancer tissues than in the adjacent normal tissues. Furthermore, there was a correlation between HMGB1 expression with TNM staging and the postoperative survival of lung cancer patients, which was similar to the results of previous studies [17]. Taken together, the results suggested that HMGB1 plays an important role in tumor progression and may be useful in the prognoses of NSCLC patients.

It is well-known that tumor cell migration and invasion are major factors leading to cancer deterioration that affects the prognoses of patients. Many studies have reported that HMGB1 is associated with cancer cell migration and invasion. Xiao and Liu suggested that silencing of HMGB1 inhibited lung cancer migration and invasion [39]. Consistent with the above studies, the Transwell results in the present study showed that pretreatment with HMGB1 improved the invasiveness of lung cancer cells.

MMPs are important in angiogenesis, growth, and metastasis of tumors [40]. As a MMP superfamily member, MMP2 is localized in a proteolytically active form on the surface of invasive cells, which has been found to play an important role in tumor invasion and metastasis [41, 42]. Activating MMP2 generates type IV collagenase, which can degrade the proteins of the extracellular matrix of tumor cells, leading to tumor cell invasion and metastasis [43].

The p65 preprotein, a transcription factor involved in cellular signal transduction, is associated with the activity of transcription factors via various mechanisms [19, 43, 44]. Previous reports suggested that the interaction between HMGB1 and RAGE triggered the activation of NF-κB, MAPK, and MMP-2/MMP-9 signaling pathways, which are associated with tumor growth, invasion, and metastasis [20, 26, 45–47]. Fujioka et al. reported a positive correlation between the transcription factor p65 and tumor metastasis [48]. Furthermore, it has been confirmed that NF-κB can activate the MMP2 promoter and activate the membrane type protease (MT1 MMP), which further leads to hydrolysis and activation of MMP2 [49, 50]. Importantly, our study showed that HMGB1 specifically promoted MMP2 and NF-κB expression. However, p65 siRNA treatment significantly reduced MMP2 expression, suggesting that HMGB1 accelerated the MMP2 expression via the NF-κB pathway to promote lung cancer migration and invasion.

Taken together, our studies showed that the high expression of HMGB1 in lung cancer, which may be used in the

prognoses of lung cancer patients, promoted lung cancer invasion and metastasis by upregulating the expression and activity of MMP-2 via an NF-κB-dependent pathway. These findings may assist in clinical diagnoses and suggest therapeutic strategies for patients with NSCLC.

Disclosure

Xiaojin Wu and Weitao Wang are co-first authors.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This study was supported by grants from the Foundation of Jiangsu Provincial Commission of Health and Family Planning (QMRC2016363); Key Talents of Medical Science in Jiangsu Province (Z201627); Xuzhou Administration of Science & Technology (KC14SH009); the 333 high-level talents of Jiangsu Province ((2016)-1903).

References

[1] C. Zhang, S. Ge, C. Hu, N. Yang, and J. Zhang, "MiRNA-218, a new regulator of HMGB1, suppresses cell migration and invasion in non-small cell lung cancer," *Acta Biochimica et Biophysica Sinica*, vol. 45, no. 12, pp. 1055–1061, 2004.

[2] X. Wu, Y. Mi, H. Yang, A. Hu, Q. Zhang, and C. Shang, "The activation of HMGB1 as a progression factor on inflammation response in normal human bronchial epithelial cells through RAGE/JNK/NF-κB pathway," *Molecular and Cellular Biochemistry*, vol. 380, no. 1-2, pp. 249–257, 2013.

[3] W. W. B. Goh, M. Fan, H. S. Low, M. Sergot, and L. Wong, "Enhancing the utility of Proteomics Signature Profiling (PSP) with Pathway Derived Subnets (PDSs), performance analysis and specialised ontologies," *BMC Genomics*, vol. 14, no. 1, article no. 35, 2013.

[4] P. Chunhacha and P. Chanvorachote, "Roles of caveolin-1 on anoikis resistance in non small cell lung cancer," *International Journal of Physiology, Pathophysiology and Pharmacology*, vol. 4, no. 3, pp. 149–155, 2012.

[5] C. Hou, H. Zhao, W. Li, and S. Cai, "Hydrogen peroxide induces high mobility group box 1 release in human bronchial epithelial cells," *Journal of Southern Medical University*, vol. 32, no. 8, pp. 1131–1134, 2012.

[6] X. Yang M and H. Yang, "Expression of high mobility group box-1 in the lung tissue and serum of patients with pulmonary tuberculosis," *Zhonghua jie he he hu xi za zhi*, vol. 36, no. 7, pp. 497–500, 2013.

[7] L. Conti, S. Lanzardo, M. Arigoni et al., "The noninflammatory role of high mobility group box 1/toll-like receptor 2 axis in the self-renewal of mammary cancer stem cells," *The FASEB Journal*, vol. 27, no. 12, pp. 4731–4744, 2013.

[8] D. Shrimali, M. K. Shanmugam, A. P. Kumar et al., "Targeted abrogation of diverse signal transduction cascades by emodin for the treatment of inflammatory disorders and cancer," *Cancer Letters*, vol. 341, no. 2, pp. 139–149, 2013.

[9] S.-Y. Chiou, Y.-S. Lee, M.-J. Jeng, P.-C. Tsao, and W.-J. Soong, "Moderate hypothermia attenuates oxidative stress injuries in alveolar epithelial A549 cells," *Experimental Lung Research*, vol. 39, no. 6, pp. 217–228, 2013.

[10] J. E. Ellerman, C. K. Brown, M. de Vera et al., "Masquerader: high mobility group box-1 and cancer," *Clinical Cancer Research*, vol. 13, no. 10, pp. 2836–2848, 2007.

[11] P.-L. Liu, J.-R. Tsai, J.-J. Hwang et al., "High-mobility group box 1-mediated matrix metalloproteinase-9 expression in non-small cell lung cancer contributes to tumor cell invasiveness," *American Journal of Respiratory Cell and Molecular Biology*, vol. 43, no. 5, pp. 530–538, 2010.

[12] J.-L. Wang, D.-W. Wu, Z.-Z. Cheng, W.-Z. Han, S.-W. Xu, and N.-N. Sun, "Expression of High Mobility Group Box-B1 (HMGB-1) and Matrix metalloproteinase-9 (MMP-9) in non-small cell lung cancer (NSCLC)," *Asian Pacific Journal of Cancer Prevention*, vol. 15, no. 12, pp. 4865–4869, 2014.

[13] D. Süren, M. Yildirim, Ö. Demirpençe et al., "The role of High Mobility Group Box 1 (HMGB1) in colorectal cancer," *Medical Science Monitor*, vol. 20, pp. 530–537, 2014.

[14] Y. R. Choi, H. Kim, H. J. Kang et al., "Overexpression of high mobility group box 1 in gastrointestinal stromal tumors with KIT mutation," *Cancer Research*, vol. 63, no. 9, pp. 2188–2193, 2003.

[15] Q. Hao, X.-Q. Du, X. Fu, and J. Tian, "Expression and clinical significance of HMGB1 and RAGE in cervical squamous cell carcinoma," *Zhonghua Zhong Liu Za Zhi*, vol. 30, no. 4, pp. 292–295, 2008.

[16] Y. Li, J. Tian, X. Fu et al., "Serum high mobility group box protein 1 as a clinical marker for ovarian cancer," *Neoplasma*, vol. 61, no. 5, pp. 579–584, 2014.

[17] A. Feng, Z. Tu, and B. Yin, "The effect of HMGB1 on the clinicopathological and prognostic features of non-small cell lung cancer," *Oncotarget*, vol. 7, pp. 20507–20519, 2016.

[18] K. Jakubowska, W. Naumnik, W. Niklińska, and E. Chyczewska, "Clinical significance of HMGB-1 and TGF-β level in serum and BALF of advanced non-small cell lung cancer," *Advances in Experimental Medicine and Biology*, vol. 852, pp. 49–58, 2015.

[19] X. Zhang, H. Wang, and J. Wang, "Expression of HMGB1 and NF-?B p65 and its significance in non-small cell lung cancer," *Wspolczesna Onkologia*, vol. 17, no. 4, pp. 350–355, 2013.

[20] G.-H. Shang, C.-Q. Jia, H. Tian et al., "Serum high mobility group box protein 1 as a clinical marker for non-small cell lung cancer," *Respiratory Medicine*, vol. 103, no. 12, pp. 1949–1953, 2009.

[21] W. Naumnik, W. Nilklińska, M. Ossolińska, and E. Chyczewska, "Serum levels of HMGB1, survivin, and VEGF in patients with advanced non-small cell lung cancer during chemotherapy," *Folia Histochemica et Cytobiologica*, vol. 47, no. 4, pp. 703–709, 2009.

[22] X. Shen, L. Hong, H. Sun, M. Shi, and Y. Song, "The expression of high-mobility group protein box 1 correlates with the progression of non-small cell lung cancer," *Oncology Reports*, vol. 22, no. 3, pp. 535–539, 2009.

[23] K.-K. Sun, C. Ji, X. Li et al., "Overexpression of high mobility group protein B1 correlates with the proliferation and metastasis of lung adenocarcinoma cells," *Molecular Medicine Reports*, vol. 7, no. 5, pp. 1678–1682, 2013.

[24] Q. Xia, J. Xu, H. Chen et al., "Association between an elevated level of HMGBI and non-small-cell lung cancer: A meta-analysis and literature review," *OncoTargets and Therapy*, vol. 9, pp. 3917–3923, 2016.

[25] H. J. Huttunen, C. Fages, and H. Rauvala, "Receptor for advanced glycation end products (RAGE)-mediated neurite outgrowth and activation of NF-κB require the cytoplasmic domain of the receptor but different downstream signaling pathways," *The Journal of Biological Chemistry*, vol. 274, no. 28, pp. 19919–19924, 1999.

[26] A. Taguchi, D. C. Blood, G. Del Toro et al., "Blockade of RAGE-amphoterin signalling suppresses tumour growth and metastases," *Nature*, vol. 405, no. 6784, pp. 354–360, 2000.

[27] J. S. Park, F. Gamboni-Robertson, Q. He et al., "High mobility group box 1 protein interacts with multiple Toll-like receptors," *American Journal of Physiology-Cell Physiology*, vol. 290, no. 3, pp. C917–C924, 2006.

[28] B. Schmalfeldt, D. Prechtel, K. Härting et al., "Increased expression of matrix metalloproteinases (MMP)-2, MMP-9, and the urokinase-type plasminogen activator is associated with progression from benign to advanced ovarian cancer," *Clinical Cancer Research*, vol. 7, no. 8, pp. 2396–2404, 2001.

[29] T. Koshiba, R. Hosotani, M. Wada et al., "Involvement of matrix metalloproteinase-2 activity in invasion and metastasis of pancreatic carcinoma," *Cancer*, vol. 82, no. 4, pp. 642–650, 1998.

[30] K. M. Panth, T. Van Den Beucken, R. Biemans et al., "In vivo optical imaging of MMP2 immuno protein antibody: Tumor uptake is associated with MMP2 activity," *Scientific Reports*, vol. 6, Article ID 22198, 2016.

[31] M. Karin, "Nuclear factor-κB in cancer development and progression," *Nature*, vol. 441, no. 7092, pp. 431–436, 2006.

[32] J. Fucikova, P. Kralikova, A. Fialova et al., "Human tumor cells killed by anthracyclines induce a tumor-specific immune response," *Cancer Research*, vol. 71, no. 14, pp. 4821–4833, 2011.

[33] C.-J. Yang, Y.-P. Liu, H.-Y. Dai et al., "Nuclear HDAC6 inhibits invasion by suppressing NF-kB/MMP2 and is inversely correlated with metastasis of non-small cell lung cancer," *Oncotarget* , vol. 6, no. 30, pp. 30263–30276, 2015.

[34] E. Sabattini, K. Bisgaard, S. Ascani et al., "The EnVision(TM)+ system: A new immunohistochemical method for diagnostics and research. Critical comparison with the APAAP, ChemMate(TM), CSA, LABC, and SABC techniques," *Journal of Clinical Pathology*, vol. 51, no. 7, pp. 506–511, 1998.

[35] Y. M. Kim et al., "Endostatin inhibits endothelial and tumor cellular invasion by blocking the activation and catalytic activity of matrix metalloproteinase," *Cancer Res*, vol. 60, pp. 5410–5413, 2000.

[36] M. Toth, A. Sohail, and R. Fridman, "Assessment of Gelatinases (MMP-2 and MMP-9) by Gelatin Zymography," in *Metastasis Research Protocols*, M. Dwek, S. A. Brooks, and U. Schumacher, Eds., vol. 878, pp. 121–135, Humana Press, 2012.

[37] Y. Zhang, J. Dong, P. He et al., "Genistein inhibit cytokines or growth factor-induced proliferation and transformation phenotype in fibroblast-like synoviocytes of rheumatoid arthritis," *Inflammation*, vol. 35, no. 1, pp. 377–387, 2012.

[38] L. L. Mantell, W. R. Parrish, and L. Ulloa, "HMGB-1 as a therapeutic target for infectious and inflammatory disorders," *Shock*, vol. 25, no. 1, pp. 4–11, 2006.

[39] P. Xiao and W. L. Liu, "MiR-142-3p functions as a potential tumor suppressor directly targeting HMGB1 in non-small-cell lung carcinoma," *International Journal of Clinical and Experimental Pathology*, vol. 8, no. 9, pp. 10800–10807, 2015.

[40] T. Klein and R. Bischoff, "Physiology and pathophysiology of matrix metalloproteases," *Amino Acids*, vol. 41, no. 2, pp. 271–290, 2011.

[41] U. B. Hofmann, J. R. Westphal, G. N. P. Van Muijen, and D. J. Ruiter, "Matrix metalloproteinases in human melanoma," *Journal of Investigative Dermatology*, vol. 115, no. 3, pp. 337–344, 2000.

[42] J. C. T. Wong, S. K. Chan, D. F. Schaeffer et al., "Absence of MMP2 expression correlates with poor clinical outcomes in rectal cancer, and is distinct from MMP1-related outcomes in colon cancer," *Clinical Cancer Research*, vol. 17, no. 12, pp. 4167–4176, 2011.

[43] A. Jacob, J. Jing, J. Lee et al., "Rab40b regulates trafficking of MMP2 and MMP9 during invadopodia formation and invasion of breast cancer cells," *Journal of Cell Science*, vol. 126, no. 20, pp. 4647–4658, 2013.

[44] Q. Ruan and Y. H. Chen, "Nuclear factor-κB in immunity and inflammation: The Treg and Th17 connection," *Advances in Experimental Medicine and Biology*, vol. 946, pp. 207–221, 2012.

[45] I. E. Dumitriu, P. Baruah, A. A. Manfredi, M. E. Bianchi, and P. Rovere-Querini, "HMGB1: guiding immunity from within," *Trends in Immunology*, vol. 26, no. 7, pp. 381–387, 2005.

[46] H. Wang, H. Yang, and K. J. Tracey, "Extracellular role of HMGB1 in inflammation and sepsis," *Journal of Internal Medicine*, vol. 255, no. 3, pp. 320–331, 2004.

[47] M. Takada, K. Hirata, T. Ajiki, Y. Suzuki, and Y. Kuroda, "Expression of Receptor for Advanced Glycation End products (RAGE) and MMP-9 in human pancreatic cancer cells," *Hepato-Gastroenterology*, vol. 51, no. 58, pp. 928–930, 2004.

[48] S. Fujioka, G. M. Sclabas, C. Schmidt et al., "Function of nuclear factor kappaB in pancreatic cancer metastasis," *Clin Cancer Res*, vol. 9, pp. 346–354, 2003.

[49] L. Yang, G. L. Shi, C. X. Song, and S. F. Xu, "Relationship between genetic polymorphism of MCP-1 and non-small-cell lung cancer in the han nationality of North China," *Genetics and Molecular Research*, vol. 9, no. 2, pp. 765–771, 2010.

[50] H. Sato, T. Takino, Y. Okada et al., "A matrix metalloproteinase expressed on the surface of invasive tumour cells," *Nature*, vol. 370, no. 6484, pp. 61–65, 1994.

Hindawi
BioMed Research International
Volume 2018, Article ID 8406857, 9 pages
https://doi.org/10.1155/2018/8406857

Research Article

Computational Approach to Investigating Key GO Terms and KEGG Pathways Associated with CNV

YuanYuan Luo,[1] Yan Yan,[1] Shiqi Zhang,[2] and Zhen Li[1]

[1]*Department of Ophthalmology, School of Medicine, Renji Hospital, Shanghai Jiao Tong University, Shanghai 200127, China*
[2]*Department of Biostatistics, University of Copenhagen, Copenhagen, Denmark*

Correspondence should be addressed to Zhen Li; lizhen1981_1@126.com

Received 26 December 2017; Revised 28 February 2018; Accepted 6 March 2018; Published 11 April 2018

Academic Editor: Jialiang Yang

Choroidal neovascularization (CNV) is a severe eye disease that leads to blindness, especially in the elderly population. Various endogenous and exogenous regulatory factors promote its pathogenesis. However, the detailed molecular biological mechanisms of CNV have not been fully revealed. In this study, by using advanced computational tools, a number of key gene ontology (GO) terms and KEGG pathways were selected for CNV. A total of 29 validated genes associated with CNV and 17,639 nonvalidated genes were encoded based on the features derived from the GO terms and KEGG pathways by using the enrichment theory. The widely accepted feature selection method—maximum relevance and minimum redundancy (mRMR)—was applied to analyze and rank the features. An extensive literature review for the top 45 ranking features was conducted to confirm their close associations with CNV. Identifying the molecular biological mechanisms of CNV as described by the GO terms and KEGG pathways may contribute to improving the understanding of the pathogenesis of CNV.

1. Introduction

Choroidal neovascularization (CNV) is a serious eye disease involving the abnormal growth of blood vessels in the choroid region [1–3]. The growth originates from a break in Bruch's membrane, and subsequently the new blood vessels penetrate into the subretinal pigment epithelium [4]. From an epidemiological perspective, CNV is a major cause of pathological visual loss in aging populations [5]. Clinically, age-related macular degeneration (ARMD), myopia, and presumed ocular histoplasmosis syndrome (POHS) are the three major pathogeneses attributed to CNV [6–8]. The Wisconsin Beaver Dam Eye Study [9] confirmed that up to 90% of visual loss in ARMD is secondary to CNV. Given that ARMD is the most common cause of visual loss in people older than 50 years, CNV is speculated to be directly linked to such pathological visual loss. Aside from ARMD, the other two pathological processes—myopia [10] and POHS [11]—are also linked to pathological visual loss. This finding validates the specific role of CNV in pathological visual loss.

Clinically, most patients with CNV share a group of characteristic signs and symptoms, including painless loss of vision, metamorphopsia, paracentral or central scotoma, and apparent changes in image size perception [12, 13]. Generally, patients with these complaints need further physical examinations on blood, fluid, lipid exudation, and retinal pigment epithelial detachment for accurate diagnoses [14]. However, for the final differential diagnosis, laboratory tests are the golden criteria. Generally, the laboratory studies for CNV involve three main techniques, namely, fluorescein angiography [15], indocyanine green angiography [16], and spectral domain optical coherence tomography [17]. For the patients with confirmed diagnosis, due to the unclear pathological mechanisms of CNV, anti-VEGF treatment that counters angiogenesis is the only preferred clinical therapeutic approach [17]. However, the injection burden limits the long-term application of such anti-VEGF treatment. Therefore, more detailed pathological mechanisms of CNV need to be revealed to promote the development and application of new drugs against the disease.

Recent publications have partially revealed the detailed pathological mechanisms of CNV, which involve the interactions between genetic factors and exogenous environments. For the environmental factors, the personal physical factors

induced by the exogenous factors are directly involved in the pathogenesis [18]. Age, obesity, high cholesterol, and high blood pressure aggravate the progression of CNV and further contribute to the occurrence of complications [19, 20]. Aside from these so-called physical exogenous factors, various genetic factors are also connected to the initiation and progression of CNV. Given that CNV is a highly specific disease with an abnormal angiogenesis, genes associated with angiogenesis, such as VEGF [21] and FGF2 [22], definitely participate in the pathological processes, which have been widely confirmed by reliable experiments. In addition to these genes, a specific gene called CFI participates in CNV and induces gradual visual loss and myopia; this finding is based on the sequencing data of CNV families [23]. Furthermore, a specific study [24] on the East Asian population with 2119 patients and 5691 controls revealed a group of effective hereditary and sporadic virulence genes that participate in CNV, mapping out the detailed genetic blueprint of CNV. Some trials were also conducted in the bioinformatics field. Zhang et al. [25] presented a specific computational routine for the identification of CNV-associated genes, indicating the efficacy and accuracy of computational application in such field.

As mentioned earlier, the genetic basis and the environmental influences of CNV have been revealed. However, its biological molecular mechanisms have not been explained thoroughly. Here, the detailed biological processes, cellular components, and molecular functions that may participate in the pathogenesis of CNV were screened out by using computational methods. In this study, GO [26] and KEGG [26, 27] pathways were introduced as two effective bioinformatics tools to accurately describe such items [27]. Based on widely known biological processes associated with CNV, an effective network was rebuilt, and novel biological processes described by the GO and KEGG items were screened out. Recent publications have validated these highly correlative biological processes, thus supporting the efficacy and accuracy of our prediction. With the use of computational methods, a group of functional biological processes that may participate in the potential pathogenesis of CNV were screened out, and for the first time, the detailed pathological mechanisms of CNV were described at the level of comprehensive biological processes instead of genes. The results contributed to the understanding of the development and progression of CNV.

2. Materials and Methods

This study aimed to extract some key GO terms and KEGG pathways that share close biological associations with CNV by using a computational framework. The flowchart of our method is illustrated in Figure 1 for the easy understanding of this work.

2.1. Materials. In 2012, Newman et al. [28] reported a number of genes that are related to AMD. We downloaded the "Additional file 3" in their study [28], in which genes associated with AMD in literature either by genetic linkage or as expression biomarker were listed. Since CNV was a subtype of AMD, we further filtered the genes. Only the genes in

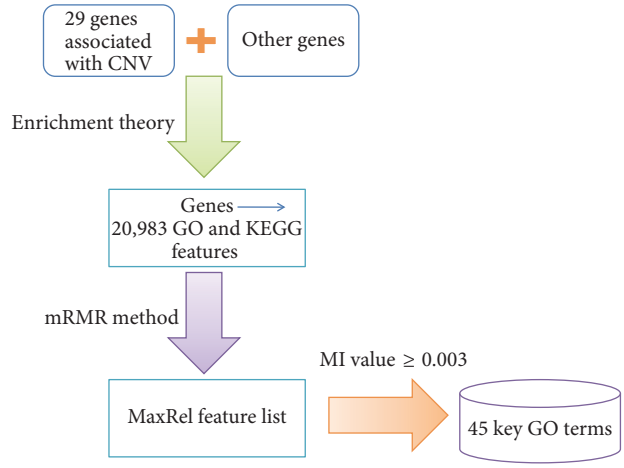


Figure 1: Flowchart of selecting the key GO terms and KEGG pathways related to CNV.

CNV Up or CNV Down modules from "Additional file 5" in Newman et al.'s study were kept and at last, 35 CNV genes were obtained. CNV Up or CNV Down modules were generated by network clustering of differentially expressed genes with a permuted $p < 0.1$ and fold change $\geqslant 1.5$ among 31 normal, 7 MD1, 4 MD2, 17 Dry AMD, 2 GA, 4 CNV, and 3 GA/CNV samples. Therefore, the final CNV genes we used were both reported by literature and differentially expressed.

The obtained 35 CNV genes were mapped onto their Ensembl IDs. We excluded IDs that are not in the PPI network reported in STRING (Version 10.0) [29]. 38 Ensembl IDs were accessed. The GO terms and KEGG pathways were used to investigate the difference between CNV-related genes and others; thus, the Ensembl IDs without a GO term and KEGG pathway information were excluded. A total of 29 Ensembl IDs were left. These IDs were the positive samples in this study. The other 17,639 Ensembl IDs were the negative samples and comprised the dataset together with the positive samples in this study. The genes belonging to the positive and negative samples are provided in Supplementary Material S1.

2.2. Feature Vector. The goal of this study was to refine important GO terms and KEGG pathways that are associated with CNV genes. To fulfill that goal, all the genes in the dataset were needed to be represented by all the GO terms and KEGG pathways. Here, the enrichment theory [30] of the GO term and KEGG pathway was used to transform the genes into numeric values, which indicated the biological relationships between the genes and GO terms (KEGG pathways). Comparing with the direct binary annotation of whether a gene has a specific GO term or KEGG pathway, the score obtained by the enrichment theory can indicate the significance of overlap between a gene set and a GO or KEGG function in the genome background. It is more robust than the binary qualitative measurement [31]. To date, this theory has been widely applied to investigate different gene- or protein-related problems [30, 32–41]. After each gene was represented by a larger number of features, by applying a

feature selection method described in Section 2.3, the key GO term or KEGG pathway features were extracted to distinguish the difference between the positive and negative samples. The encoding procedure follows.

GO Enrichment Score. The GO enrichment score was utilized to represent the association between a GO term and an involved gene as a numeric value. For a given GO term, such as GO_j, and a gene g, the gene set G_1 consisted of genes annotated to GO_j and gene set G_2 consisted of the neighbor genes of g in the protein–protein interaction network reported in STRING (http://string-db.org/) [29], a well-organized database providing known and predicted protein–protein interactions. On the basis of the preceding items, the GO enrichment score of GO_j and g can be defined as the $-\log_{10}$ of the hypergeometric test p value [30, 32–35] of G_1 and G_2 according to the following equation:

$$\mathrm{ES}_{\mathrm{GO}}\left(g, \mathrm{GO}_j\right) = -\log_{10}\left(\sum_{k=m}^{n} \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}\right), \quad (1)$$

where N is the total number of genes in humans, M is the number of genes in G_1, n is the number of genes in G_2, and m is the number of the common genes of G_1 and G_2. A large enrichment score of GO_j and g indicated a close relationship between them. In this study, 20,686 GO terms were considered. Thus, 20,686 GO enrichment scores were calculated for each gene in the dataset, which were obtained by using an in-house program using R function phyper. The R code is "score ← −log10(phyper(numWdrawn − 1, numW, numB, numDrawn, lower.tail = FALSE))," where numW, numB, and numDrawn correspond to the number of genes annotated to GO_j, number of genes not annotated to GO_j, and number of neighbor genes g, respectively.

KEGG Enrichment Score. Similar to the GO enrichment score, the KEGG pathway score was calculated using the same theory to represent the quantitative associations between the KEGG pathways and genes in the dataset. For a given KEGG pathway K_j and a gene g, G_1 was a gene set containing genes in K_j and G_2 had the same meaning as described in preceding paragraph. The KEGG enrichment score shared a similar definition with the GO enrichment score between K_j and g, which was formulated as

$$\mathrm{ES}_{\mathrm{KEGG}}\left(g, K_j\right) = -\log_{10}\left(\sum_{k=m}^{n} \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}\right), \quad (2)$$

where N, M, n, and m share similar definitions as described in (1). In addition, a high score yielded by a KEGG pathway K_j and a gene g indicated their strong associations. Here, 297 KEGG pathways were considered and resulted in 297 KEGG enrichment scores for each gene, which were also obtained by using an in-house program using R function phyper.

Accordingly, each gene in the dataset was encoded by a combination of 20,686 GO term and 297 KEGG pathway features and was defined as a feature vector with a total of 20,983 elements:

$$f(g) = \left(\mathrm{ES}_{\mathrm{GO}}(g, \mathrm{GO}_1), \ldots, \mathrm{ES}_{\mathrm{GO}}(g, \mathrm{GO}_{20686}), \mathrm{ES}_{\mathrm{KEGG}}(g, K_1), \ldots, \mathrm{ES}_{\mathrm{KEGG}}(g, K_{291})\right)^T. \quad (3)$$

2.3. Feature Selection. As described in Section 2.2, each gene in the dataset was encoded with 20,983 features derived from the GO terms and KEGG pathways. Some of them shared closer biological associations with CNV. Thus, advanced tools were necessary to extract these important features that played essential roles in the development of CNV. Here, a reliable and widely accepted feature selection method, namely, maximum relevance and minimum redundancy (mRMR) [42], was adopted to analyze all 20,983 features. The mRMR method, proposed by Peng et al. [42], is a useful tool to analyze the feature space of complicated biological problems. To date, many investigations related to complicated biological systems or problems have applied this method to analyze their feature space and extract important information [34, 36, 43–52].

In the mRMR method, two excellent criteria were proposed to rank the features: (1) maximum relevance and (2) minimum redundancy. According to their names, the former criterion measures the importance of features by relying on their correlation to target variable, whereas the latter criterion provides a guarantee that the selected features also have minimum redundancies. If one decides to construct an optimal feature subspace, both maximum relevance and minimum redundancy should be used. In this study, the purpose was to extract important features that are closely related to CNV rather than construct an optimal feature subspace. Therefore, only the criterion of maximum relevance was employed to rank the features in this study. The maximum relevance of each feature was measured by the mutual information (MI) between the feature and the target variable. For each feature, f was a variable representing the values in all samples and c was the target variable. The MI was calculated as follows:

$$I(c, f) = \iint p(c, f) \log \frac{p(c, f)}{p(c)\, p(f)} dc\, df, \quad (4)$$

where $p(c)$ and $p(f)$ are the marginal probabilities of c and f and $p(c, f)$ is their joint probabilistic distribution. According to (4), MI measures the mutual dependence between two variables.

Based on the MI value assigned to each feature, the feature ranking list called MaxRel feature list was obtained and formulated as follows:

$$F = [f_1, f_2, \ldots, f_N], \quad (5)$$

where N is the total number of features in the feature space. A high rank received by a feature indicates a strong association with CNV. Based on the properties of the top ranked features, a new insight into the CNV can be proposed for the investigation of the corresponding GO terms and KEGG pathways.

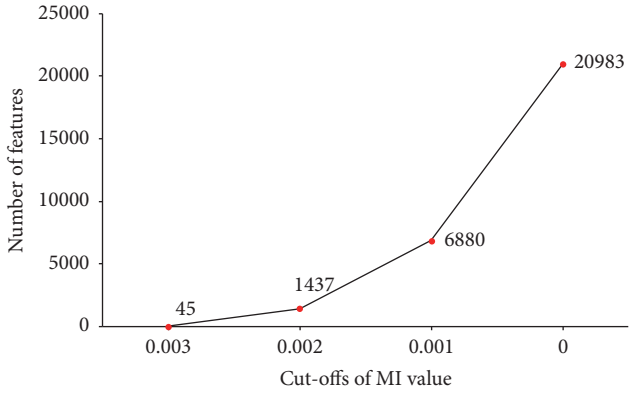


Figure 2: The number of selected features under different cut-offs of MI values.

3. Results and Discussion

3.1. Results. As described in Section 2.2, a total of 20,983 GO terms and KEGG pathway features were encoded in each gene in dataset. Then, according to their relevance to the target variables, these features were ranked in the descending order by using the maximum relevance criterion described in Section 2.3. The output feature list, called MaxRel feature list, was built and obtained (Supplementary Material S2).

As mentioned in the preceding paragraphs, not all GO terms or KEGG pathways shared equal roles on influencing the progression of CNV. Thus, extracting the key GO terms or KEGG pathways was necessary. By applying the maximum relevance criterion, all the features were ranked by their relevance to the target variables, and the rank of a corresponding feature in the output MaxRel feature list for a GO term or KEGG pathway indicated its association with CNV. According to their MI values, some GO terms or KEGG pathways received high MI values in the MaxRel feature list; these features were extracted and their importance was further investigated. To determine the cut-off of MI value, a curve was plotted in Figure 2, which shows the number of selected features under different cut-offs of MI value. It can be observed that the cut-off 0.003 was a proper choice, resulting in 45 features. These features would be given further literature review. Their detailed information is listed in Table 1. All the 45 features corresponded to important GO terms. The following section provides a detailed discussion on these GO terms.

3.2. Analysis of Key GO Terms. As mentioned earlier, based on our current computational methods, a group of functional biological processes that may directly contribute to the initiation and progression of CNV as a pathological mechanism were screened out. In the prediction list, the top 45 biological processes described by the GO terms as optimal CNV-associated biological processes were selected. Due to the limitation of such manuscript, an individual analysis of all the items was not feasible. Therefore, the top terms were chosen, and their respective connection with CNV according to recent publications was discussed. According to recent publications, such GO terms can be summarized into three major subgroups: angiogenesis, local neural metabolism, and immune-associated biological processes. The detailed discussion follows.

3.2.1. Analysis of Angiogenesis Associated Biological Processes. The two GO terms in our prediction list—*GO: 0031091* and *GO: 0031093*—both describe the functional cellular components of platelet alpha-granules. In 2015, a specific study [53] on proangiogenic responses confirmed that the release of platelet alpha-granules promotes angiogenesis. No direct connections were revealed between platelet alpha-granules and CNV; however, abnormal angiogenesis plays an irreplaceable role and may be the core pathological biological process during the initiation and progression of CNV [54]. Therefore, GO items associated with platelet alpha-granules, such as cellular components GO: 0031091 and GO: 0031093, are definitely associated with CNV. This result validated the efficacy and accuracy of our prediction.

GO term *GO: 0038133* describes a detailed pathway called the ERBB2-ERBB3 signaling pathway. According to recent publications, this signaling pathway contributes to the regulation of cell survival and tumorigenesis [55, 56]. As for the detailed connections between the ERBB2-ERBB3 signaling pathway and CNV, mediated by miR-199a and miR-125b, ERBB2 and ERBB3 as two functional components of our predicted biological process have been confirmed to contribute to the regulation of vascular endothelial growth factor secretion and the stimulation of angiogenesis in multiple tissues, including the eyes [57–59]. Given the core initiative functions of angiogenesis for CNV, the predicted biological process called the ERBB2-ERBB3 signaling pathway is a potential CNV-associated GO term. Moreover, the next predicted GO term, called *GO: 0038129*, also describes the ERBB3-associated signaling pathway. This finding not only implied the prediction consistency of the current computational methods but also further confirmed the specific role of such pathways during the initiation and progression of CNV.

GO: 0031983 was the next predicted GO term and describes the vesicle lumen as a functional cellular component. As the parent term of GO: 0060205 describing the cytoplasmic vesicle lumen, such cellular component definitely is associated with the initiation and progression of CNV. As for detailed literature evidence, in 2009, a specific study on the vascular permeability and pathological angiogenesis of CNV confirmed that the vesicle lumen in living cells is related to the vascular hyperpermeability and abnormal angiogenesis [60]. Vascular hyperpermeability [61] and abnormal angiogenesis are both specific symptoms of CNV [62]; thus, such biological processes are potential CNV-associated biological processes.

The next GO term, called *GO: 0005576*, describes a general term called extracellular region. Various extracellular substances participate in the pathogenesis of CNV, including LOX [63], LOXL2 [63], Thy-1 [64], and integrins [65]. Such specific substances may play irreplaceable roles during the initiation and progression of CNV; thus, this GO term that describes the extracellular regions of a certain focus is a potential CNV-associated biological process. *GO: 0035767*, as the next predicted GO in our prediction list, describes an effective biological process called endothelial cell chemotaxis.

Table 1: Top 45 key GO terms associated with CNV.

GO term ID	GO term	GO description	MI value	Rank
GO: 0031091	Platelet alpha-granule	Cellular component	0.003	1
GO: 0031093	Platelet alpha-granule lumen	Cellular component	0.003	2
GO: 0060205	Cytoplasmic membrane-bounded vesicle lumen	Cellular component	0.003	3
GO: 0038133	ERBB2-ERBB3 signaling pathway	Biological process	0.003	4
GO: 0038129	ERBB3 signaling pathway	Biological process	0.003	5
GO: 1902847	Regulation of neuronal signal transduction	Biological process	0.003	6
GO: 0061517	Macrophage proliferation	Biological process	0.003	7
GO: 1902949	Positive regulation of tau protein kinase activity	Biological process	0.003	8
GO: 0061518	Microglial cell proliferation	Biological process	0.003	9
GO: 0031983	Vesicle lumen	Cellular component	0.003	10
GO: 0005576	Extracellular region	Cellular component	0.003	11
GO: 0035767	Endothelial cell chemotaxis	Biological process	0.003	12
GO: 0002580	Regulation of antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	Biological process	0.003	13
GO: 0044421	Extracellular region part	Cellular component	0.003	14
GO: 0007603	Phototransduction, visible light	Biological process	0.003	15
GO: 0001948	Glycoprotein binding	Molecular function	0.003	16
GO: 0072562	Blood microparticle	Cellular component	0.003	17
GO: 0044650	Adhesion of symbiont to host cell	Biological process	0.003	18
GO: 0019062	Virion attachment to host cell	Biological process	0.003	19
GO: 0010466	Negative regulation of peptidase activity	Biological process	0.003	20
GO: 0009584	Detection of visible light	Biological process	0.003	21
GO: 0010951	Negative regulation of endopeptidase activity	Biological process	0.003	22
GO: 0001654	Eye development	Biological process	0.003	23
GO: 0002581	Negative regulation of antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	Biological process	0.003	24
GO: 0052547	Regulation of peptidase activity	Biological process	0.003	25
GO: 0052548	Regulation of endopeptidase activity	Biological process	0.003	26
GO: 0005791	Rough endoplasmic reticulum	Cellular component	0.003	27
GO: 0050839	Cell adhesion molecule binding	Molecular function	0.003	28
GO: 0071307	Cellular response to vitamin K	Biological process	0.003	29
GO: 1902430	Negative regulation of beta-amyloid formation	Biological process	0.003	30
GO: 0005604	Basement membrane	Cellular component	0.003	31
GO: 0030023	Extracellular matrix constituent conferring elasticity	Molecular function	0.003	32
GO: 2000768	Positive regulation of nephron tubule epithelial cell differentiation	Biological process	0.003	33
GO: 1903002	Positive regulation of lipid transport across blood brain barrier	Biological process	0.003	34
GO: 1903000	Regulation of lipid transport across blood brain barrier	Biological process	0.003	35
GO: 1903001	Negative regulation of lipid transport across blood brain barrier	Biological process	0.003	36
GO: 1902951	Negative regulation of dendritic spine maintenance	Biological process	0.003	37
GO: 1902999	Negative regulation of phospholipid efflux	Biological process	0.003	38
GO: 1901627	Negative regulation of postsynaptic membrane organization	Biological process	0.003	39
GO: 2001139	Negative regulation of phospholipid transport	Biological process	0.003	40
GO: 0046911	Metal chelating activity	Molecular function	0.003	41
GO: 0030574	Collagen catabolic process	Biological process	0.003	42
GO: 0007423	Sensory organ development	Biological process	0.003	43
GO: 0001968	Fibronectin binding	Molecular function	0.003	44
GO: 0051346	Negative regulation of hydrolase activity	Biological process	0.003	45

Based on recent publications, such biological process is involved in the activation of platelets [66] and exosome-mediated antiangiogenesis [67]. Platelet activation [68] and angiogenesis [69] are directly connected to the initiation and progression of CNV; therefore, this predicted GO term is quite significant in the pathogenesis of CNV.

3.2.2. Analysis of Local Neural Metabolism Associated Biological Processes. *GO: 0060205*, as another cellular component associated item, describes the cytoplasmic vesicle lumen. Based on recent publications, such cellular component participates in autophagy and secretion-associated biological processes in living cells [70, 71]. As for the biological connections between the cytoplasmic vesicle lumen and CNV, the predicted GO-associated biological processes, such as autophagy and substance secretion, have all been confirmed to be involved in the initiation and progression of CNV [72, 73]. This result implied the accuracy and efficacy of our prediction. *GO: 1902847* describes the regulation of neuronal signal transduction. In the biological process of neuronal signal transduction, a specific gene called IKK2 has been confirmed to be significant [74]. Coincidentally, the inhibition of IKK2 has been widely used in the treatment against CNV, indicating the specific role of IKK2 during the pathogenesis of CNV [75, 76]. Therefore, connected by such functional gene IKK2, the predicted biological processes associated with neuronal signal transduction may also be related to CNV. This finding validates the efficacy and accuracy of our prediction. As the next predicted GO, *GO: 1902949* describes the positive regulation of tau protein kinase activity. Tau protein is a major pathological factor that contributes to the initiation and progression of Alzheimer's disease (AD) [77–79]. During the initiation and progression of AD, another specific protein called apolipoprotein E4 (apoE4) interacts with our predicted tau protein [80] and participates in the pathogenesis of AD [81]. Given that recent studies also validated the specific role of apoE4 in neovascularization [82] and its potential functions in CNV [23], tau protein associated kinase activity is reasonably connected to CNV-associated biological characteristics.

3.2.3. Analysis of Immune-Associated Biological Processes. The GO term *GO: 0061517* describes the proliferation of a specific immune-associated cell subgroup: macrophage. Based on recent publications, macrophages contribute to CNV by regulating CCR2-dependent and proangiogenic biological processes [83, 84], indicating that the proliferation of such gene is definitely related to the progression of the disease [85]. Apart from the proliferation of macrophage, the proliferation of another effective cell subgroup called microglial cells is also predicted to contribute to CNV by *GO: 0061518* in our prediction list. Mediated by neuroprotectin D1, microglial ramifications and redistribution participate in the pathological processes of CNV [86]. Therefore, as a functional neuronal cell subtype with specific microglial ramifications [86], this predicted GO is reasonably connected to CNV [87]. Besides these predicted biological processes, several functional GOs in the top 45 predicted GO terms have been reported to participate in CNV-associated biological processes. These functional GO terms include GO: 0002580 (regulation of antigen processing and presentation of peptide or polysaccharide antigen via MHC class II) [88], GO: 0044421 (extracellular region) [89], and GO: 0007603 (phototransduction, visible light) [90, 91]. These results confirmed the efficacy and accuracy of our prediction.

4. Conclusion

Based on our presented computational method, a group of functional biological functions that have been confirmed by recent publications to be related to the pathogenesis of CNV were screened out. Such predicted biological processes not only further revealed the detailed pathological mechanisms of CNV but also provided a new tool to identify potential functional disease-associated biological processes in multiple categories of the disease. Finally, we will try our best to develop a computational method based on some extracted features in this study to predict novel CNV genes in the future.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Authors' Contributions

YuanYuan Luo and Yan Yan contributed equally to this work.

Supplementary Materials

The supplementary materials consist of two files. Supplementary Material S1: positive and negative samples in the dataset represented by their gene symbols. Supplementary Material S2: MaxRel feature list as an output of the mRMR method with ranked 20,983 features. *(Supplementary Materials)*

References

[1] F. B. Pereira, C. E. Veloso, G. T. Kokame, and M. B. Nehemy, "Characteristics of Neovascular Age-Related Macular Degeneration in Brazilian Patients," *Ophthalmologica*, vol. 234, no. 4, pp. 233–242, 2015.

[2] D. Karagiannis, G. A. Kontadakis, K. Kaprinis et al., "Treatment of myopic choroidal neovascularization with intravitreal ranibizumab injections: The role of age," *Clinical Ophthalmology*, vol. 11, pp. 1197–1201, 2017.

[3] C. M. G. Cheung, J. J. Arnold, F. G. Holz et al., "Myopic choroidal neovascularization: review, guidance, and consensus statement on management," *Ophthalmology*, 2017.

[4] J. J. Jung, C. Y. Chen, S. Mrejen et al., "The incidence of neovascular subtypes in newly diagnosed neovascular age-related macular degeneration," *American Journal of Ophthalmology*, vol. 158, no. 4, pp. 769–779.e2, 2014.

[5] K. Neelam, C. M. G. Cheung, K. Ohno-Matsui, T. Y. Y. Lai, and T. Y. Wong, "Choroidal neovascularization in pathological myopia," *Progress in Retinal and Eye Research*, vol. 31, no. 5, pp. 495–525, 2012.

[6] D. P. Han, J. T. McAllister, D. V. Weinberg, J. E. Kim, and W. J. Wirostko, "Combined intravitreal anti-VEGF and verteporfin

photodynamic therapy for juxtafoveal and extrafoveal choroidal neovascularization as an alternative to laser photocoagulation," *Eye*, vol. 24, no. 4, pp. 713–716, 2010.

[7] M. El Mellaoui, A. El Ouafi, Z. El Hansali, A. Bouzidi, S. Iferkhas, and A. Laktaoui, "Presumed ocular histoplasmosis syndrome," *Journal Français d'Ophtalmologie*, vol. 38, no. 9, pp. 892-893, 2015.

[8] M. Laghmari and O. Lezrek, "Presumed ocular histoplasmosis syndrome (POHS)," *Pan African Medical Journal*, vol. 18, article no. 268, p. 268, 2014.

[9] R. Klein, B. E. K. Klein, and K. L. P. Linton, "Prevalence of age-related maculopathy. The Beaver Dam Eye Study," *Ophthalmology*, vol. 99, no. 6, pp. 933–943, 1992.

[10] I. M. Ghafour, D. Allan, and W. S. Foulds, "Common causes of blindness and visual handicap in the west of Scotland," *British Journal of Ophthalmology*, vol. 67, no. 4, pp. 209–213, 1983.

[11] S. S. Feman, S. F. Podgorski, and M. K. Penn, "Blindness from presumed ocular histoplasmosis in Tennessee," *Ophthalmology*, vol. 89, no. 12, pp. 1295–1298, 1982.

[12] K. Hatz and C. Prünte, "Polypoidal choroidal vasculopathy in Caucasian patients with presumed neovascular age-related macular degeneration and poor ranibizumab response," *British Journal of Ophthalmology*, vol. 98, no. 2, pp. 188–194, 2014.

[13] C. M. G. Cheung, B. K. Loh, X. Li et al., "Choroidal thickness and risk characteristics of eyes with myopic choroidal neovascularization," *Acta Ophthalmologica*, vol. 91, no. 7, pp. e580–e581, 2013.

[14] M. A. Bonini Filho, T. E. de Carlo, D. Ferrara et al., "Association of choroidal neovascularization and central serous chorioretinopathy with optical coherence tomography angiography," *JAMA Ophthalmology*, vol. 133, no. 8, pp. 899–906, 2015.

[15] W. Liu, H. Li, R. S. Shah et al., "Simultaneous optical coherence tomography angiography and fluorescein angiography in rodents with normal retina and laser-induced choroidal neovascularization," *Optics Expresss*, vol. 40, no. 24, pp. 5782–5785, 2015.

[16] A. Kawamura, M. Yuzawa, R. Mori, M. Haruyama, and K. Tanaka, "Indocyanine green angiographic and optical coherence tomographic findings support classification of polypoidal choroidal vasculopathy into two types," *Acta Ophthalmologica*, vol. 91, no. 6, pp. e474–e481, 2013.

[17] P. Iacono, M. Battaglia Parodi, A. Papayannis et al., "Fluorescein angiography and spectral-domain optical coherence tomography for monitoring Anti-VEGF therapy in myopic choroidal neovascularization," *Ophthalmic Research*, vol. 52, no. 1, pp. 25–31, 2014.

[18] I. C. Munch, A. Linneberg, and M. Larsen, "Precursors of age-related macular degeneration: associations with physical activity, obesity, and serum lipids in the Inter99 Eye Study," *Investigative Ophthalmology & Visual Science*, vol. 54, no. 6, pp. 3932–3940, 2013.

[19] C.-C. Hsu, S.-J. Chen, A.-F. Li, and F.-L. Lee, "Systolic blood pressure, choroidal thickness, and axial length in patients with myopic maculopathy," *Journal of the Chinese Medical Association*, vol. 77, no. 9, pp. 487–491, 2014.

[20] X. Zhang, M. Li, F. Wen et al., "Different impact of high-density lipoprotein-related genetic variants on polypoidal choroidal vasculopathy and neovascular age-related macular degeneration in a Chinese Han population," *Experimental Eye Research*, vol. 108, pp. 16–22, 2013.

[21] P. S. Muether, I. Neuhann, C. Buhl, M. M. Hermann, B. Kirchhof, and S. Fauser, "Intraocular growth factors and cytokines in patients with dry and neovascular age-related macular degeneration," *Retina*, vol. 33, no. 9, pp. 1809–1814, 2013.

[22] J. R. O. De Dias, E. B. Rodrigues, M. Maia, O. Magalhat~es Jr., F. M. Penha, and M. E. Farah, "Cytokines in neovascular age-related macular degeneration: Fundamentals of targeted combination therapy," *British Journal of Ophthalmology*, vol. 95, no. 12, pp. 1631–1637, 2011.

[23] N. Leveziel, Y. Yu, R. Reynolds et al., "Genetic factors for choroidal neovascularization associated with high myopia," *Investigative Ophthalmology & Visual Science*, vol. 53, no. 8, pp. 5004–5009, 2012.

[24] C. Y. Cheng, K. Yamashiro, and L. J. Chen, "New loci and coding variants confer risk for age-related macular degeneration in East Asians," *Nature Communications*, vol. 6, p. 6063, 2015.

[25] J. Zhang, Y. Suo, Y.-H. Zhang et al., "Mining for genes related to choroidal neovascularization based on the shortest path algorithm and protein interaction information," *Biochimica et Biophysica Acta (BBA) - General Subjects*, vol. 1860, no. 11, pp. 2740–2749, 2016.

[26] The Gene Ontology Consortium, "Gene ontology consortium: going forward," *Nucleic Acids Research*, vol. 43, no. 1, pp. D1049–D1056, 2015.

[27] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "KEGG as a reference resource for gene and protein annotation," *Nucleic Acids Research*, vol. 44, no. 1, pp. D457–D462, 2016.

[28] A. M. Newman, N. B. Gallo, L. S. Hancox et al., "Systems-level analysis of age-related macular degeneration reveals global biomarkers and phenotype-specific functional networks," *Genome Medicine*, vol. 4, no. 2, article 16, 2012.

[29] D. Szklarczyk, A. Franceschini, S. Wyder et al., "STRING v10: protein-protein interaction networks, integrated over the tree of life," *Nucleic Acids Research*, vol. 43, pp. D447–D452, 2015.

[30] P. Carmona-Saez, M. Chagoyen, F. Tirado, J. M. Carazo, and A. Pascual-Montano, "GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists," *Genome Biology*, vol. 8, no. 1, article R3, 2007.

[31] T. Huang, C. Wang, G. Zhang, L. Xie, and Y. Li, "SySAP: a system-level predictor of deleterious single amino acid polymorphisms," *Protein & Cell*, vol. 3, no. 1, pp. 38–43, 2012.

[32] T. Huang, L. Chen, Y. Cai, and K. Chou, "Classification and analysis of regulatory pathways using graph property, biochemical and physicochemical property, and functional property," *PLoS ONE*, vol. 6, no. 9, Article ID e25297, 2011.

[33] T. Huang, J. Zhang, Z. Xu et al., "Deciphering the effects of gene deletion on yeast longevity using network and machine learning approaches," *Biochimie*, vol. 94, no. 4, pp. 1017–1025, 2012.

[34] L. Chen, Y.-H. Zhang, G. Lu, T. Huang, and Y.-D. Cai, "Analysis of cancer-related lncRNAs using gene ontology and KEGG pathways," *Artificial Intelligence in Medicine*, vol. 76, pp. 27–36, 2017.

[35] L. Chen, Y.-H. Zhang, S. Wang, Y. Zhang, T. Huang, and Y.-D. Cai, "Prediction and analysis of essential genes using the enrichments of gene ontology and KEGG pathways," *PLoS ONE*, vol. 12, no. 9, Article ID e0184129, 2017.

[36] L. Chen, Y. Zhang, M. Zheng, T. Huang, and Y. Cai, "Identification of compound–protein interactions through the analysis of gene ontology, KEGG enrichment for proteins and molecular fragments of compounds," *Molecular Genetics and Genomics*, vol. 291, no. 6, pp. 2065–2079, 2016.

[37] J. Li, L. Chen, S. Wang et al., "A computational method using the random walk with restart algorithm for identifying novel epigenetic factors," *Molecular Genetics and Genomics*, vol. 293, no. 1, pp. 293–301, 2018.

[38] S. Lu, Y. Yan, Z. Li et al., "Determination of genes related to uveitis by utilization of the random walk with restart algorithm on a protein–protein interaction network," *International Journal of Molecular Sciences*, vol. 18, no. 5, article no. 1045, 2017.

[39] L. Chen, J. Yang, Z. Xing et al., "An integrated method for the identification of novel genes related to oral cancer," *PLoS ONE*, vol. 12, no. 4, p. e0175185, 2017.

[40] Y. Zhou, B. Li, Y. Zhang, L. Chen, and X. Kong, "Feature classification and analysis of lung cancer related genes through gene ontology and KEGG pathways," *Current Bioinformatics*, vol. 11, no. 1, pp. 40–50, 2016.

[41] J. Yang, L. Chen, X. Kong, T. Huang, and Y. D. Cai, "Analysis of tumor suppressor genes based on gene ontology and the KEGG pathway," *PLoS ONE*, vol. 9, no. 9, Article ID e107202, 2014.

[42] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[43] S. Wang, Y. Zhang, J. Lu, W. Cui, J. Hu, and Y. Cai, "Analysis and identification of aptamer-compound interactions with a maximum relevance minimum redundancy and nearest neighbor algorithm," *BioMed Research International*, vol. 2016, Article ID 8351204, 9 pages, 2016.

[44] L. Chen, C. Chu, T. Huang, X. Kong, and Y. Cai, "Prediction and analysis of cell-penetrating peptides using pseudo-amino acid composition and random forest models," *Amino Acids*, vol. 47, no. 7, pp. 1485–1493, 2015.

[45] T. Huang, M. Wang, and Y.-D. Cai, "Analysis of the preferences for splice codes across tissues," *Protein & Cell*, vol. 6, no. 12, pp. 904–907, 2015.

[46] Z. Li, X. Zhou, Z. Dai, and X. Zou, "Classification of G-protein coupled receptors based on support vector machine with maximum relevance minimum redundancy and genetic algorithm," *BMC Bioinformatics*, vol. 11, article 325, 2010.

[47] Q. Ni and L. Chen, "A feature and algorithm selection method for improving the prediction of protein structural classes," *Combinatorial Chemistry High Throughput Screening*, vol. 20, no. 7, pp. 612–621, 2017.

[48] B.-Q. Li, L.-L. Zheng, K.-Y. Feng, L.-L. Hu, G.-H. Huang, and L. Chen, "Prediction of linear B-cell epitopes with mRMR feature selection and analysis," *Current Bioinformatics*, vol. 11, no. 1, pp. 22–31, 2016.

[49] H. Mohabatkar, M. M. Beigi, K. Abdolahi, and S. Mohsenzadeh, "Prediction of allergenic proteins by means of the concept of Chou's pseudo amino acid composition and a machine learning approach," *Medicinal Chemistry*, vol. 9, no. 1, pp. 133–137, 2013.

[50] L. Chen, S. Wang, Y. Zhang et al., "Identify Key Sequence Features to Improve CRISPR sgRNA Efficacy," *IEEE Access*, vol. 5, pp. 26582–26590, 2017.

[51] L. Chen, C. Chu, and K. Feng, "Predicting the types of metabolic pathway of compounds using molecular fragments and sequential minimal optimization," *Combinatorial Chemistry & High Throughput Screening*, vol. 19, no. 2, pp. 136–143, 2016.

[52] L. Chen, Y.-H. Zhang, G. Huang et al., "Discriminating cirRNAs from other lncRNAs using a hierarchical extreme learning machine (H-ELM) algorithm with feature selection," *Molecular Genetics and Genomics*, vol. 293, no. 1, pp. 137–149, 2018.

[53] J. Etulain, H. A. Mena, S. Negrotto, and M. Schattner, "Stimulation of PAR-1 or PAR-4 promotes similar pattern of VEGF and endostatin release and pro-angiogenic responses mediated by human platelets," *Platelets*, vol. 26, no. 8, pp. 799–804, 2015.

[54] H. Xu, F. Zeng, D. Shi, X. Sun, X. Chen, and Y. Bai, "Focal choroidal excavation complicated by choroidal neovascularization," *Ophthalmology*, vol. 121, no. 1, pp. 246–250, 2014.

[55] C. F. McDonagh, A. Huhalov, B. D. Harms et al., "Antitumor activity of a novel bispecific antibody that targets the ErbB2/ErbB3 oncogenic unit and inhibits heregulin-induced activation of ErbB3," *Molecular Cancer Therapeutics*, vol. 11, no. 3, pp. 582–593, 2012.

[56] V. Fock, K. Plessl, P. Draxler et al., "Neuregulin-1-mediated ErbB2-ErbB3 signalling protects human trophoblasts against apoptosis to preserve differentiation," *Journal of Cell Science*, vol. 128, no. 23, pp. 4306–4316, 2015.

[57] J. He, Y. Jing, W. Li et al., "Roles and Mechanism of miR-199a and miR-125b in Tumor Angiogenesis," *PLoS ONE*, vol. 8, no. 2, Article ID e56647, 2013.

[58] L. Yen, X. You, A. Al Moustafa et al., "Heregulin selectively upregulates vascular endothelial growth factor secretion in cancer cells and stimulates angiogenesis," *Oncogene*, vol. 19, no. 31, pp. 3460–3469, 2000.

[59] K. S. Russell, D. F. Stern, P. J. Polverini, and J. R. Bender, "Neuregulin activation of ErbB receptors in vascular endothelium leads to angiogenesis," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 277, no. 6, pp. H2205–H2211, 1999.

[60] S.-H. Chang, D. Feng, J. A. Nagy, T. E. Sciuto, A. M. Dvorak, and H. F. Dvorak, "Vascular permeability and pathological angiogenesis in caveolin-1-null mice," *The American Journal of Pathology*, vol. 175, no. 4, pp. 1768–1776, 2009.

[61] P. Daull, C. A. Paterson, B. D. Kuppermann, and J.-S. Garrigue, "A preliminary evaluation of dexamethasone palmitate emulsion: A novel intravitreal sustained delivery of corticosteroid for treatment of macular Edema," *Journal of Ocular Pharmacology and Therapeutics*, vol. 29, no. 2, pp. 258–269, 2013.

[62] H. Du, X. Sun, M. Guma et al., "JNK inhibition reduces apoptosis and neovascularization in a murine model of age-related macular degeneration," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 110, no. 6, pp. 2377–2382, 2013.

[63] T. V. Bergen, R. Spangler, D. Marshall et al., "The role of LOX and LOXL2 in the pathogenesis of an experimental model of choroidal neovascularization," *Investigative Ophthalmology & Visual Science*, vol. 56, no. 9, pp. 5280–5289, 2015.

[64] H. Wang, X. Han, E. Kunz, and M. Elizabeth Hartnett, "Thy-1 regulates VEGF-mediated choroidal endothelial cell activation and migration: Implications in neovascular age-related macular degeneration," *Investigative Ophthalmology & Visual Science*, vol. 57, no. 13, pp. 5525–5534, 2016.

[65] T. Nakajima, M. Hirata, T. R. Shearer, and M. Azuma, "Mechanism for laser-induced neovascularization in rat choroid: accumulation of integrin alpha chain-positive cells and their ligands," *Molecular Vision*, vol. 20, pp. 864–871, 2014.

[66] A. Janowska-Wieczorek, M. Wysoczynski, J. Kijowski et al., "Microvesicles derived from activated platelets induce metastasis and angiogenesis in lung cancer," *International Journal of Cancer*, vol. 113, no. 5, pp. 752–760, 2005.

[67] A. R. Hajrasouliha, G. Jiang, Q. Lu et al., "Exosomes from retinal astrocytes contain antiangiogenic components that inhibit

laser-induced choroidal neovascularization," *The Journal of Biological Chemistry*, vol. 288, no. 39, pp. 28058–28067, 2013.

[68] K. Birke, E. Lipo, M. T. Birke, and R. Kumar-Singh, "Topical Application of PPADS Inhibits Complement Activation and Choroidal Neovascularization in a Model of Age-Related Macular Degeneration," *PLoS ONE*, vol. 8, no. 10, Article ID e76766, 2013.

[69] J. H. Qi, Q. Ebrahem, M. Ali et al., "Tissue Inhibitor of Metalloproteinases-3 Peptides Inhibit Angiogenesis and Choroidal Neovascularization in Mice," *PLoS ONE*, vol. 8, no. 3, Article ID e55667, 2013.

[70] M. Zhang, S. J. Kenny, L. Ge, K. Xu, and R. Schekman, "Translocation of interleukin-1beta into a vesicle intermediate in autophagy-mediated secretion," *eLife*, vol. 4, 2015.

[71] V. Malhotra and P. Erlmann, "The Pathway of Collagen Secretion," *Annual Review of Cell and Developmental Biology*, vol. 31, pp. 109–124, 2015.

[72] A. Klettner, A. Kauppinen, J. Blasiak, J. Roider, A. Salminen, and K. Kaarniranta, "Cellular and molecular mechanisms of age-related macular degeneration: from impaired autophagy to neovascularization," *The International Journal of Biochemistry & Cell Biology*, vol. 45, no. 7, pp. 1457–1497, 2013.

[73] R. Zhang, Z. Liu, H. Zhang, Y. Zhang, and D. Lin, "The COX-2-selective antagonist (NS-398) inhibits choroidal neovascularization and subretinal fibrosis," *PLoS ONE*, vol. 11, no. 1, Article ID e0146808, 2016.

[74] E. Tsaousidou, L. Paeger, B. F. Belgardt et al., "Distinct Roles for JNK and IKK Activation in Agouti-Related Peptide Neurons in the Development of Obesity and Insulin Resistance," *Cell Reports*, vol. 9, no. 4, pp. 1495–1506, 2014.

[75] S. Gaddipati, Q. Lu, R. B. Kasetti et al., "IKK2 inhibition using TPCA-1-Loaded PLGA microparticles attenuates laser-induced choroidal neovascularization and macrophage recruitment," *PLoS ONE*, vol. 10, no. 3, Article ID e0121185, 2015.

[76] H. Lu, Q. Lu, S. Gaddipati et al., "IKK2 inhibition attenuates laser-induced choroidal neovascularization," *PLoS ONE*, vol. 9, no. 1, Article ID e87530, 2014.

[77] S. Crunkhorn, "Antisense oligonucleotide reverses tau pathology," *Nature Reviews Drug Discovery*, vol. 16, no. 3, pp. 166-166, 2017.

[78] T. Uchihara, K. Endo, H. Kondo et al., "Tau pathology in aged cynomolgus monkeys is progressive supranuclear palsy/corticobasal degeneration- but not Alzheimer disease-like -Ultrastructural mapping of tau by EDX," *Acta Neuropathologica Communications*, vol. 4, no. 1, p. 118, 2016.

[79] H. Malkki, "Alzheimer disease: BACE1 inhibition could block CSF tau increase," *Nature Reviews Neurology*, vol. 13, no. 1, p. 6, 2017.

[80] O. Liraz, A. Boehm-Cagan, and D. M. Michaelson, "ApoE4 induces Aβ42, tau, and neuronal pathology in the hippocampus of young targeted replacement apoE4 mice," *Molecular Neurodegeneration*, vol. 8, no. 1, article no. 16, 2013.

[81] J.-T. Yu, L. Tan, and J. Hardy, "Apolipoprotein e in Alzheimer's disease: An update," *Annual Review of Neuroscience*, vol. 37, pp. 79–100, 2014.

[82] R. Antes, S. Salomon-Zimri, S. C. Beck et al., "VEGF mediates apoE4-induced neovascularization and synaptic pathology in the choroid and retina," *Current Alzheimer Research*, vol. 12, no. 4, pp. 323–334, 2015.

[83] T. A. Krause, A. F. Alex, D. R. Engel, C. Kurts, and N. Eter, "VEGF-production by CCR2-dependent macrophages contributes to laser-induced choroidal neovascularization," *PLoS ONE*, vol. 9, no. 4, Article ID e94313, 2014.

[84] S. Horie, S. J. Robbie, J. Liu et al., "CD200R signaling inhibits pro-angiogenic gene expression by macrophages and suppresses choroidal neovascularization," *Scientific Reports*, vol. 3, article no. 3072, 2013.

[85] L. He and A. G. Marneros, "Doxycycline inhibits polarization of macrophages to the proangiogenic M2-type and subsequent neovascularization," *The Journal of Biological Chemistry*, vol. 289, no. 12, pp. 8019–8028, 2014.

[86] K. G. Sheets, B. Jun, Y. Zhou, and etal., "Microglial ramification and redistribution concomitant with the attenuation of choroidal neovascularization by neuroprotectin D1," *Molecular Vision*, vol. 19, pp. 1747–1759, 2013.

[87] M.-C. Carrasco, J. Navascués, M. A. Cuadros et al., "Migration and ramification of microglia in quail embryo retina organotypic cultures," *Developmental Neurobiology*, vol. 71, no. 4, pp. 296–315, 2011.

[88] P. L. Penfold, J. G. Wong, J. Gyory, and F. A. Billson, "Effects of triamcinolone acetonide on microglial morphology and quantitative expression of MHC-II in exudative age-related macular degeneration," *Clinical & Experimental Ophthalmology*, vol. 29, no. 3, pp. 188–192, 2001.

[89] S. Binder, B. V. Stanzel, I. Krebs, and C. Glittenberg, "Transplantation of the RPE in AMD," *Progress in Retinal and Eye Research*, vol. 26, no. 5, pp. 516–554, 2007.

[90] V. P. Papastefanou, V. Nogueira, G. Hay et al., "Choroidal naevi complicated by choroidal neovascular membrane and outer retinal tubulation," *British Journal of Ophthalmology*, vol. 97, no. 8, pp. 1014–1019, 2013.

[91] M. B. Parod, P. Iacono, and F. Bandello, "Correspondence of leakage on fluorescein angiography and optical coherence tomography parameters in diagnosis and monitoring of myopic choroidal neovascularization treated with bevacizumab," *Retina*, vol. 36, no. 1, pp. 104–109, 2016.

Hindawi
BioMed Research International
Volume 2018, Article ID 3145689, 8 pages
https://doi.org/10.1155/2018/3145689

Research Article

AEG-1 Contributes to Metastasis in Hypoxia-Related Ovarian Cancer by Modulating the HIF-1alpha/NF-kappaB/VEGF Pathway

Xiaoyu Yu, Yan Wang, Huilei Qiu, Hongtao Song, Di Feng, Yang Jiang, Shuzhe Deng, Hongxue Meng, and Jingshu Geng

Department of Pathology, Harbin Medical University Cancer Hospital, Harbin, China

Correspondence should be addressed to Jingshu Geng; gengjingshu@yeah.net

Received 1 December 2017; Accepted 11 January 2018; Published 25 March 2018

Academic Editor: Jialiang Yang

Objective. Ovarian carcinoma represents one of the deadliest malignancies among female cancer patients. Astrocyte-elevated gene-1 (AEG-1) participates in the ontogenesis of multiple human malignant diseases. Here we evaluated AEG-1, hypoxia-inducible factor- (HIF-) 1α, nuclear factor kappa-light-chain-enhancer of activated B cells (NF-κB), and vascular endothelial growth factor (VEGF) amounts in hypoxia induced ovarian carcinoma cells. This study aimed to explore the mechanism by which AEG-1 regulates metastasis in hypoxia induced ovarian carcinoma. *Patients and Methods*. AEG-1, HIF-1α, and VEGF protein amounts were evaluated by immunohistochemistry in 40 and 170 normal ovary and ovarian cancer tissue specimens, respectively. In addition, AEG-1, HIF-1α, NF-κB, and VEGF mRNA and protein levels were determined by reverse quantified RT-PCR and WB, respectively, at different time periods (0–24 h) in epithelial ovarian cancer (EOC) SKOV3 cells treated in a hypoxia incubator. Furthermore, NF-κB and VEGF gene and protein expression levels in AEG-1 knockdown EOC cells were quantitated by RT-PCR and WB, respectively. *Results*. AEG-1, HIF-1α, and VEGF amounts were significantly elevated in EOC tissue samples compared with normal ovary specimens ($p < 0.001$). Positive expression of HIF-1α and AEG-1 was associated with higher metastatic rate ($p < 0.01$), lower FIGO stage ($p < 0.001$), and degree of differentiation ($p < 0.001$). Meanwhile, EOC SKOV3 cells grew upon exposure to hypoxia for 8 h ($p < 0.001$); at this time point, AEG-1, HIF-1α, NF-κB, and VEGF amounts peaked ($p < 0.001$), at both the gene and the protein levels. After AEG-1 knockdown, HIF-1α, NF-κB, and VEGF amounts were significantly decreased in EOC SKOV3 cells, also under hypoxic conditions ($p < 0.01$). *Conclusions*. As an independent prognostic factor, AEG-1 was found to be significantly associated with hypoxia in ovarian cancer by regulating the HIF-1alpha/NF-kappaB/VEGF pathway. Therefore, AEG-1 may be useful in determining disease stage and prognosis in ovarian cancer.

1. Introduction

Ovarian carcinoma, a commonly encountered primary malignancy, represents one of the deadliest malignant tumors in female patients around the world. Astrocyte-elevated gene-1 (AEG-1) plays multiple roles and acts as an important molecule regulating a variety of events in carcinogenesis. Mounting evidence indicates that AEG-1 expression is elevated in a wide range of malignancies [1, 2], such as hepatocellular, gallbladder, renal cell, breast, lung, and ovarian carcinomas [3]. The above works showed that AEG-1 constitutes a critical transcription factor in cancer metastasis and invasion. However, the mechanism by which AEG-1 regulates metastasis in ovarian carcinoma remains largely unknown.

HIF-1α, firstly identified as a transcription factor, is activated by hypoxic stress. In addition, it acts as a hypoxia-inducible nuclear factor in VSMCs and has significant functions in hypoxic responses in human cells, modulating the transcription of hypoxia-inducible genes [4].

This work evaluated AEG-1, HIF-1α, NF-κB, and VEGF protein and mRNA amounts in epithelial ovarian cancer (EOC) cells cultured under hypoxic conditions and found potential associations of AEG-1, NF-κB, and VEGF expression levels with hypoxia induced ovarian cancer growth.

Table 1: AEG-1, VEGF, and HIF-1α levels in ovarian cancer and normal tissue specimens.

Group	Cases (N)	AEG-1 high expression n (%)	p	VEGF high expression n (%)	p	HIF-1α high expression n (%)	p
EOC tissues	170	107 (62.9)	<0.001	93 (54.7)	<0.001	102 (60.0)	<0.001
Normal ovary	40	2 (5)		4 (10)		2 (5)	

2. Materials and Methods

2.1. Patient Population. The current study had approval from the Cancer Hospital of Harbin Medical University, Harbin, China. Written informed consent was obtained from each patient. Samples were collected from 170 and 40 patients with EOC and normal ovary suffering from other diseases, respectively, who underwent surgery between February 2007 and February 2009 in the Cancer Hospital of Harbin Medical University. Tumor stage in each patient was evaluated according to the International Federation of Gynecology and Obstetrics (FIGO) classification. Histopathological grades were determined according to the World Health Organization criteria.

2.2. Immunohistochemical Staining. AEG-1, HIF-1α, and VEGF protein levels were assessed immunohistochemically in biopsy samples after paraffin-embedding, by the avidin-biotin immunoperoxidase method, as instructed by the manufacturer. After dewaxing and rehydration by standard methods, the sections were incubated with primary antibodies targeting human AEG-1 (ab45338, Abcam), HIF-1α (ab16066, Abcam), and VEGF (ab155944, Abcam) overnight at 4°C, respectively, followed by incubation with biotin-conjugated secondary antibodies (Santa Cruz, USA). Negative control slides were incubated with rabbit serum in lieu of primary antibodies.

2.2.1. Cell Culture. Human EOC SKOV3 cells were purchased from the Cell Institute of Chinese Academy of Sciences, Shanghai, and grown in McCoy's 5A medium containing 10% fetal bovine serum at 37°C in a normal 5% CO2 cell culture incubator. To induce hypoxia, cell culture was performed in an anaerobic chamber containing 1% O2, 5% CO2, and 94% N2, at 37°C for 0, 2, 4, 8, 10, 12, and 24 h, respectively.

2.2.2. Lentiviral Infection. AEG-1 knockdown lentivirus was manufactured by Shanghai GenePharma Co.,

Ltd. SKOV3 cells were plated into 3.5 cm dishes (1×10^6 cells/dish) for 24 h prior to lentiviral infection at a multiplicity of infection (MOI) of 4. Infection efficiency, assessed by fluorescence microscopy detecting GFP at 48 h after infection, was >90%.

2.2.3. Quantified Real-Time PCR Assay. Cellular RNA extraction was carried out with TRIzol reagent (Takara, Otsu, Shi, Japan). Then, real-time PCR was performed on a DNA Engine Opticon™ sequence detector. Primers for AEG-1, HIF-1α, and VEGF were designed with Primer Bank. Mean Ct (triplicate experiments) was employed for data analysis, with the endogenous control U6 snRNA used for normalization.

2.2.4. Immunoblot. Total protein was obtained from SKOV3 cells using RIPA buffer supplemented with proteinase/phosphatase inhibitors (Thermo, Cambridge, MA). Equal amounts of total protein were resolved by 10 or 12% SDS-PAGE, followed by transfer onto nitrocellulose membranes (Millipore, Bedford, MA). Anti-AEG-1 (1 : 500), NF-κB (1 : 1000), HIF-1α (1 : 500), and VEGF (1 : 1000) (Abcam) primary antibodies were used for detection.

2.2.5. Transwell Invasion Assay. Cells cultured in media with or without 100 nM rapamycin were harvested after 8 h of exposure to hypoxia. Those cultured under normoxic conditions in parallel were used as the control group.

2.3. Statistics. Data are mean ± SEM. Groups were compared by one-way analysis of variance (ANOVA) with SPSS 13.0 (SPSS, USA), followed by Bonferroni post hoc tests. Unpaired Student's t-test was used to assess group pairs. A statistical significance level of 0.05 was used.

3. Results

3.1. AEG-1, HIF-1α, and VEGF Levels in Ovarian Carcinoma and Noncancerous Specimens. By immunohistochemical staining, we first observed that AEG-1 was primarily localized in the cytosol of ovarian cancer cells. Meanwhile, HIF-1α was expressed in both cytoplasmic and nuclear compartments. VEGF was mainly expressed in the cytoplasmic compartment or cell membrane (Figure 1).

Assessing the expression levels of these three proteins in tumor samples and normal specimens, we found that 62.9%, 60%, and 54.7% ovarian cancer tissue specimens were positive for AEG-1, HIF-1α, and VEGF, respectively, for only 5%, 5%, and 10% obtained in normal ovarian tissue samples, respectively (all $p < 0.001$) (Table 1).

3.2. Associations of HIF-1α and AEG-1 Levels with Clinicopathologic Features of EOC Patients. AEG-1 expression in ovarian cancer specimens was significantly associated with histological type, metastasis, FIGO stage, and residual tumor but not correlated with age (Table 2). Similarly, HIF-1α levels were associated with histological type, metastasis, clinical stage, FIGO stage, and residual tumor (Table 2). Specifically, the expression levels of AEG-1 and HIF-1α were higher in stages III/IV than in stages I/II (Table 2), and the differences reached statistical significance (Table 2).

3.3. HIF-1α, VEGF, and AEG-1 Expression Levels Are Positively Correlated in Ovarian Carcinoma. Poor prognosis in ovarian cancer results from concerted effects of multiple genes. In this

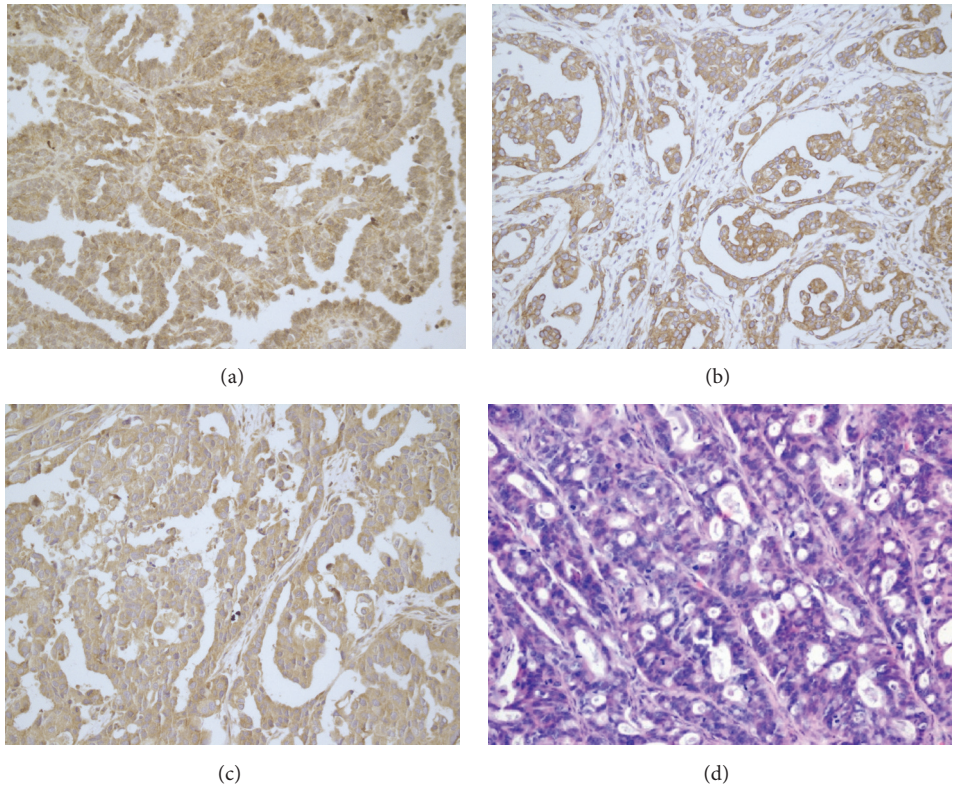


Figure 1: AEG-1, VEGF, and HIF-1α expression levels in ovarian carcinoma. (a) High AEG-1 expression; (b) high VEGF expression; (c) high HIF-1α expression; (d) H&E staining of (a)–(d) (×100).

study, AEG-1 levels were obviously associated with HIF-1α and VEGF amounts (Table 3, all $p < 0.001$). Meanwhile, HIF-1α amounts were positively associated with VEGF and AEG-1 levels (Table 3, all $p < 0.001$).

3.4. Hypoxia Increases Invasion in Ovarian Carcinoma Cells. Transwell assays were employed to estimate the invasive ability of the EOC SKOV3 cell line. The results showed that the invasive ability of SKOV3 cells was significantly enhanced in hypoxia compared with normoxia (Figures 2 and 3). These findings suggested that hypoxia increased the invasive ability of EOC cells.

Interestingly, the invasive ability of AEG-1 knockdown SKOV3 cells was overtly reduced compared with that of wildtype SKOV3 cells, in both normoxic and hypoxic culture conditions (Figures 2 and 3). This indicated that AEG-1 might be a key effector of hypoxia induced ovarian cancer growth.

3.5. Hypoxia Upregulates AEG-1, HIF-1α, NF-κB, and VEGF. As shown above, high AEG-1 expression was associated with metastasis in ovarian cancer (Table 2), and hypoxia could promote the invasive ability of cultured ovarian carcinoma cells (Figure 3). Furthermore, AEG-1 was involved in hypoxia induced ovarian cancer cell growth (Figure 3). Therefore, we hypothesized that hypoxia might affect AEG-1 expression.

SKOV3 cells were cultured under hypoxic conditions for different times (0, 2, 4, 6, 8, 10, 12, and 24 h), for total protein and RNA extraction. As expected, hypoxia treated SKOV3 cells showed markedly increased AEG-1 protein and mRNA levels, which peaked at 8 h of culture under hypoxic conditions (Figure 4, $p < 0.01$). These findings suggested that hypoxia upregulated AEG-1.

HIF-1α and VEGF levels were positively correlated with ovarian cancer metastasis, as shown above (Tables 1 and 2). Therefore, we detected level changes of HIF-1α, NF-κB, and VEGF in SKOV3 cells cultured under hypoxic conditions, and similar results were found. Indeed, HIF-1α, NF-κB, and VEGF amounts changed gradually with prolonged hypoxia (Figure 4, $p < 0.01$). Overall, our results showed that AEG-1, HIF-1α, NF-κB, and VEGF transcription levels were altered by hypoxia.

3.6. Hypoxia Associated Upregulation of HIF-1α, VEGF, and NF-κB Is Dependent on AEG-1. We found that hypoxia associated upregulation of NF-κB, HIF-1α, and VEGF in SKOV3 cells was modulated by AEG-1. Indeed, NF-κB, HIF-1α, and VEGF protein amounts were significantly decreased by AEG-1 silencing (Figure 5, $p < 0.01$). In addition, AEG-1 could regulate hypoxia induced transcription of HIF-1α, NF-κB, and VEGF in ovarian cancer cells (Figure 5).

4. Discussion

EOC is the predominant type of ovarian carcinoma and the deadliest malignancy in females, with a very high prevalence in China. The currently available diagnostic markers are not

Table 2: Associations of HIF-1α and AEG-1 levels with clinicopathologic features of EOC patients.

Variable	Cases (N)	AEG-1 high levels		HIF-1α high levels	
		n (%)	p	n (%)	p
Age			0.271		0.697
≤55	63	43 (68.3)		39 (61.9)	
>55	107	64 (59.8)		63 (58.9)	
Metastasis			0.005		0.001
Negative	28	11 (39.3)		9 (32.1)	
Positive	142	96 (67.6)		93 (65.5)	
Histological type			0.002		0.006
Serous	111	79 (71.2)		75 (67.6)	
Others	59	28 (47.5)		27 (45.8)	
Grade			0.644		0.035
G1-2	101	65 (64.4)		54 (53.5)	
G3	69	42 (60.9)		48 (69.6)	
FIGO stage			0.000		0.001
I-II	45	19 (42.2)		18 (40.0)	
III-IV	125	88 (70.4)		84 (67.2)	
Residual tumor (cm)			0.000		0.000
≤1	81	37 (45.7)		34 (42.0)	
>1	89	70 (78.7)		68 (76.4)	
Ascites (ml)			0.008		0.019
≤500	62	31 (50.0)		30 (48.4)	
>500	108	76 (70.4)		72 (66.7)	
Menopausal status			0.251		0.094
No	55	38 (69.1)		38 (69.1)	
Yes	115	69 (60.0)		64 (55.7)	
Preoperative CA125 level (U/ml)			0.913		0.820
≤35	14	9 (64.3)		8 (57.1)	
>35	156	98 (62.8)		94 (60.3)	
Chemo.			0.261		0.648
No	61	35 (57.4)		38 (62.3)	
Yes	109	72 (66.1)		64 (58.7)	

Table 3: Associations of HIF-1α, VEGF, and AEG-1 levels in EOC.

	Expression		
	AEG-1 versus VEGF	AEG-1 versus HIF-1α	VEGF versus HIF-1α
p	<0.001	<0.001	0.01

effective enough. Therefore, more adequate markers should be identified to predict metastasis and/or recurrence in EOC patients.

AEG-1 was firstly described as a gene induced in primary human fetal astrocytes. Then, studies revealed its important roles in many aspects of cancer. Clinical and functional analyses have showed that AEG-1 could be considered a potentially crucial target in the treatment of malignant neoplasms. In this study, AEG-1 was upregulated in human EOC and significantly associated with multiple clinicopathological characteristics (Table 2) [5]. These findings suggested that AEG-1 plays important roles in tumor progression and metastasis.

HIF-1α has critical functions in hypoxic response of human cells, modulating hypoxia-inducible genes [4]. It is known that HIF-1α increases blood, oxygen, and energy supply to tumors, attenuating hypoxia, [6] and has crucial functions in cancer cell metabolism, angiogenesis, and metastasis. Meanwhile, HIF-1α is regulated by many factors [7].

As shown above, 62.9%, 60.0%, and 54.7% of EOC patients had elevated AEG-1, HIF-1α, and VEGF amounts, respectively, rates which were markedly higher than control values (Table 1). These results corroborated previous findings that AEG-1 is involved in tumor progression [1, 8–20]. According to our statistical analysis, high AEG-1 and HIF-1α expression levels were commonly associated with elevated metastasis ($p < 0.01$) as well as lower FIGO stage ($p < 0.001$) and degree of differentiation ($p < 0.001$), in patients with advanced ovarian carcinoma and lymph node metastasis

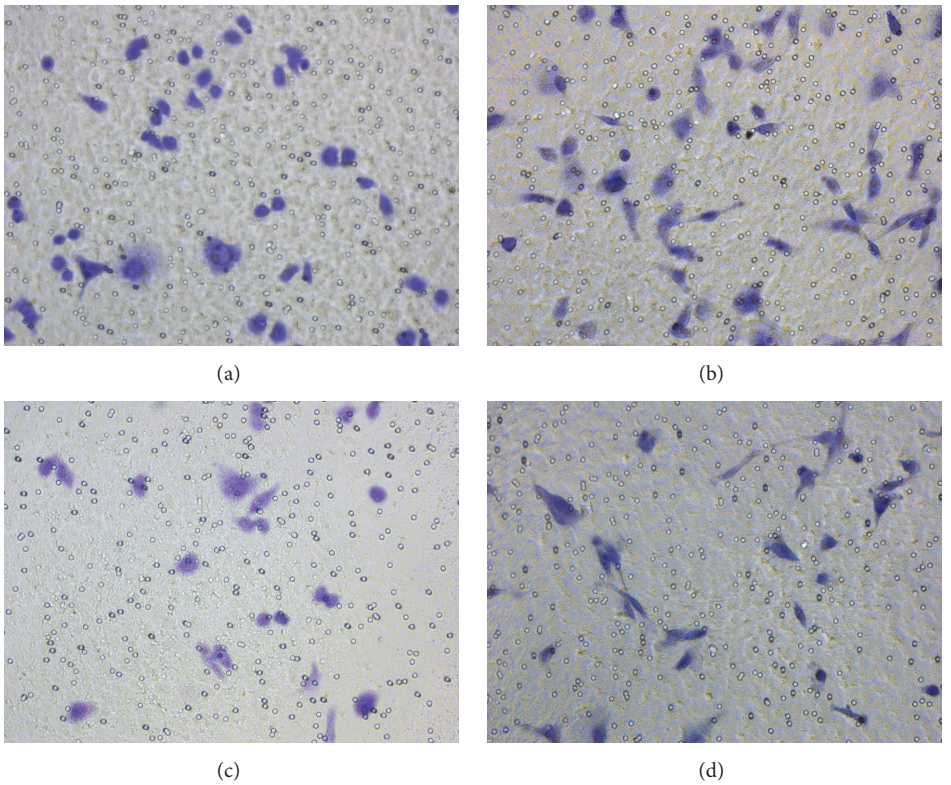


Figure 2: *Wild type and AEG-1 knockdown SKOV3 cells grown under hypoxic and normoxic conditions.* (a) SKOV3 cells in normoxia; (b) SKOV3 cells in hypoxia; (c) len-AEG-1 SKOV3 cells in normoxia; (d) len-AEG-1 SKOV3 cells in hypoxia. Data are mean ± SD from three independent experiments.

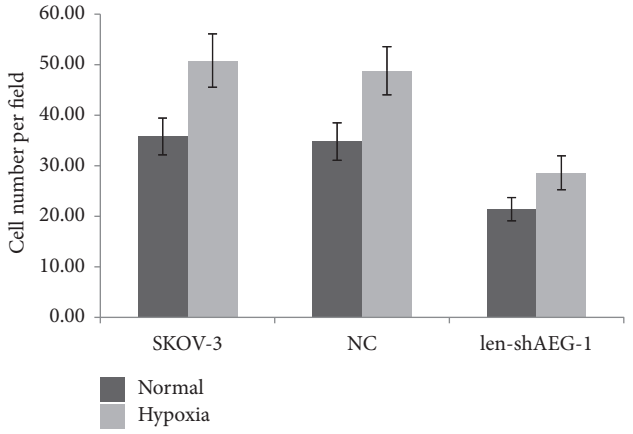


Figure 3: Hypoxia increases the invasive ability of the EOC cell line, which is reduced by AEG-1 silencing.

(Table 2), as reported in previous studies [21, 22]. This suggests that AEG-1 and HIF-1α are potential early diagnostic markers for ovarian carcinoma.

Furthermore, we demonstrated that AEG-1 levels were obviously associated with HIF-1α and VEGF amounts (Table 3, $p < 0.001$). To assess intercorrelations among these factors, AEG-1, HIF-1α, NF-κB, and VEGF mRNA and protein amounts were quantitated by quantitative real-time PCR and Western blotting, respectively, at different times in EOC cells exposed to hypoxia. Interestingly, SKOV3 cells showed significantly increased levels of the above factors after exposure to hypoxia (Figure 4), with peaks observed at 8 h ($p < 0.001$). In addition, the expression levels of NF-κB and VEGF were evaluated in AEG-1 knockdown ovarian cancer cells by RT-PCR and Western blotting, respectively. After AEG-1 knockdown, HIF-1α, NF-κB, and VEGF mRNA and protein amounts were significantly decreased in EOC SKOV3 cells cultured under hypoxic conditions (Figure 5).

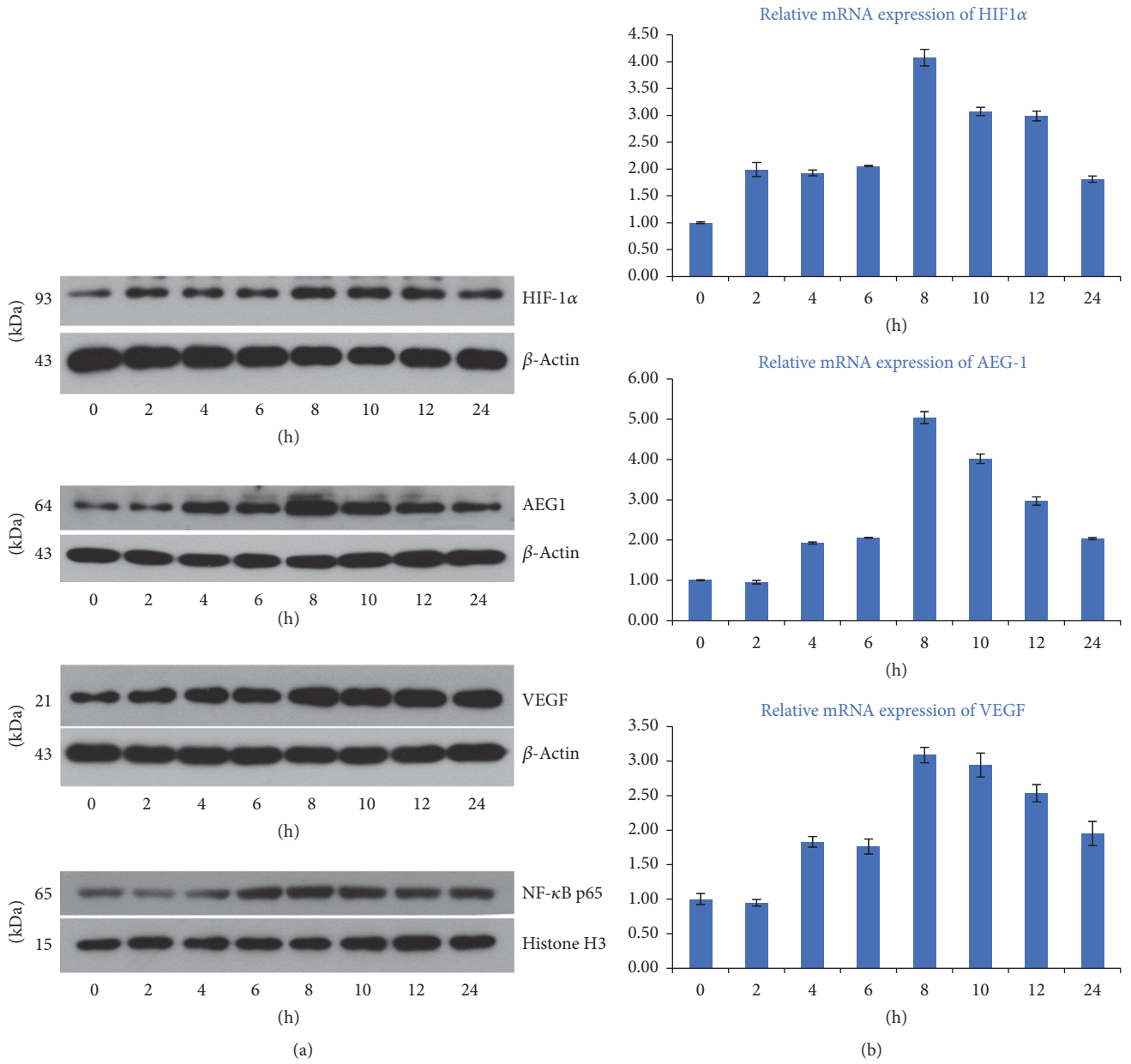


Figure 4: *AEG-1, VEGF, NF-κB, and HIF-1α expression levels in ovarian carcinoma SKOV3 cells cultured under hypoxic conditions at different time points (0, 2, 4, 8, 10, 12, and 24 h).* (a) After culture under hypoxic conditions for 8 h, AEG-1, HIF-1α, NF-κB, and VEGF protein amounts peaked ($p < 0.001$) as determined by Western blotting. (b) After culture under hypoxic conditions for 8 h, AEG-1, HIF-1α, and VEGF mRNA amounts peaked ($p < 0.001$) as determined by real-time PCR. Data are mean ± SD from three independent experiments.

As demonstrated above, HIF-1α, NF-κB, and VEGF were downregulated after AEG-1 silencing. These findings suggest that AEG-1 induces HIF-1α/NF-κB/VEGF signaling in hypoxic SKOV3 cells. In agreement, the significance of an AEG-1- dependent pathway was revealed in tumor-associated angiogenesis and cancer progression, with VEGF downstream of AEG-1 [15]. Considering the crosstalk between AEG-1 and HIF-1α/NF-κB/VEGF signaling, AEG-1 might be involved in hypoxia regulation. Consistently, tight associations of AEG-1, VEGF, and HIF-1α levels in EOC were found in this work. However, whether combining such prognostic biomarkers would improve prognosis in EOC requires further assessment.

5. Conclusions

Overall, the above findings indicated HIF-1α and AEG-1 are critical angiogenic markers and constitute potential prognostic factors in EOC exposed to a hypoxic environment. Combined expression analysis of AEG-1, HIF-1α, and VEGF may help determine the degree of malignancy, metastasis, and prognosis in EOC. Further studies and evidence are

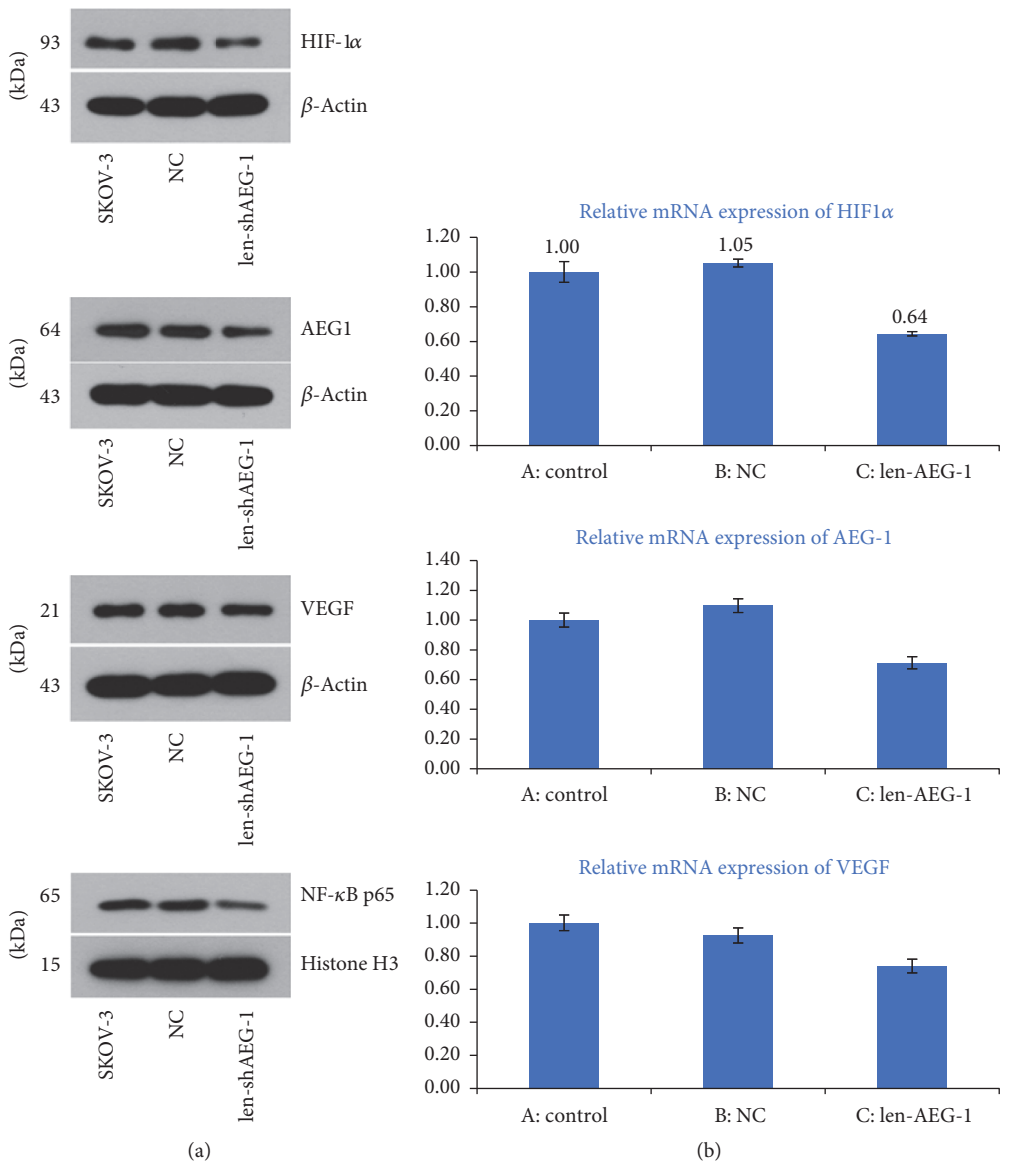


FIGURE 5: *AEG-1, HIF-1α, NF-κB, and VEGF expression levels in wild type and AEG-1 knockdown SKOV3 cells cultured under hypoxic conditions*. (a) After AEG-1 knockdown by lentiviral infection, AEG-1, HIF-1α, NF-κB, and VEGF protein amounts were significantly reduced (all $p < 0.001$) as assessed by Western blotting. (b) After AEG-1 knockdown, AEG-1, HIF-1α, and VEGF gene expression levels were significantly reduced (all $p < 0.001$) as determined by real-time PCR. Data are mean ± SD from three independent experiments.

required to evaluate whether the HIF-1α and AEG-1 proteins might help predict unfavorable biological behaviors and/or constitute targets to determine the patients benefiting from antiangiogenic agents. Additional prospective studies are warranted to confirm the current findings.

Conflicts of Interest

The authors report no conflicts of interest in this work.

Acknowledgments

This work was supported by grants from the National Natural Science Foundation of China (no. 81372785).

References

[1] D. C. Kang, Z. Z. Su, D. Sarkar, L. Emdad, D. J. Volsky, and P. B. Fisher, "Cloning and characterization of HIV-1-inducible astrocyte elevated gene-1, AEG-1," *Gene*, vol. 353, no. 1, pp. 8–15, 2005.

[2] Y. Huang and L.-P. Li, "Progress of cancer research on astrocyte elevated gene-1/Metadherin," *Oncology Letters*, vol. 8, no. 2, pp. 493–501, 2014.

[3] Z.-Z. Su, D.-C. Kang, Y. Chen et al., "Identification and cloning of human astrocyte genes displaying elevated expression after infection with HIV-1 or exposure to HIV-1 envelope glycoprotein by rapid subtraction hybridization, RaSH," *Oncogene*, vol. 21, no. 22, pp. 3592–3602, 2002.

[4] C. de Vallière, J. Cosin-Roger, S. Simmen et al., "Hypoxia Positively Regulates the Expression of pH-Sensing G-Protein–Coupled Receptor OGR1 (GPR68)," *Cellular and Molecular Gastroenterology and Hepatology*, vol. 2, no. 6, pp. 796–810, 2016.

[5] F. Meng, C. Luo, L. Ma, Y. Hu, and G. Lou, "Clinical significance of astrocyte elevated gene-1 expression in human epithelial ovarian carcinoma," *International Journal of Gynecological Pathology*, vol. 30, no. 2, pp. 145–150, 2011.

[6] S. Yasuda, S. Arii, A. Mori et al., "Hexokinase II and VEGF expression in liver tumors: correlation with hypoxia-inducible factor-1α and its significance," *Journal of Hepatology*, vol. 40, no. 1, pp. 117–123, 2004.

[7] G. L. Semenza, "Targeting HIF-1 for cancer therapy," *Nature Reviews Cancer*, vol. 3, no. 10, pp. 721–732, 2003.

[8] G. L. Semenza, "Defining the role of hypoxia-inducible factor 1 in cancer biology and therapeutics," *Oncogene*, vol. 29, no. 5, pp. 625–634, 2010.

[9] G. Hu, R. A. Chong, Q. Yang et al., "MTDH Activation by 8q22 Genomic Gain Promotes Chemoresistance and Metastasis of Poor-Prognosis Breast Cancer," *Cancer Cell*, vol. 15, no. 1, pp. 9–20, 2009.

[10] D. Sarkar, L. Emdad, S.-G. Lee, B. K. Yoo, Z.-Z. Su, and P. B. Fisher, "Astrocyte elevated gene-1: Far more than just a gene regulated in astrocytes," *Cancer Research*, vol. 69, no. 22, pp. 8529–8535, 2009.

[11] G. Hu, Y. Wei, and Y. Kang, "The multifaceted role of MTDH/AEG-1 in cancer progression," *Clinical Cancer Research*, vol. 15, no. 18, pp. 5615–5620, 2009.

[12] S.-G. Lee, Z.-Z. Su, L. Emdad, D. Sarkar, and P. B. Fisher, "Astrocyte elevated gene-1 (AEG-1) is a target gene of oncogenic Ha-ras requiring-phosphatidylinositol 3-kinase and c-Myc," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 103, no. 46, pp. 17390–17395, 2006.

[13] L. Emdad, D. Sarkar, Z.-Z. Su et al., "Corrigendum to "Astrocyte elevated gene-1: Recent insights into a novel gene involved in tumor progression, metastasis and neurodegeneration" [Pharmacol. Ther., 114 (2007) 155-170] (DOI:10.1016/j.pharmthera.2007.01.010)," *Pharmacology & Therapeutics*, vol. 115, no. 1, p. 176, 2007.

[14] D. Sarkar, S. P. Eun, L. Emdad, S.-G. Lee, Z.-Z. Su, and P. B. Fisher, "Molecular basis of nuclear factor-κB activation by astrocyte elevated gene-1," *Cancer Research*, vol. 68, no. 5, pp. 1478–1484, 2008.

[15] L. Emdad, S.-G. Lee, Z. Z. Su et al., "Astrocyte elevated gene-1 (AEG-1) functions as an oncogene and regulates angiogenesis," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 106, no. 50, pp. 21300–21305, 2009.

[16] J. Li, L. Yang, L. Song et al., "Astrocyte elevated gene-1 is a proliferation promoter in breast cancer via suppressing transcriptional factor FOXO1," *Oncogene*, vol. 28, no. 36, pp. 3188–3196, 2009.

[17] P. Su, Q. Zhang, and Q. Yang, "Immunohistochemical analysis of Metadherin in proliferative and cancerous breast tissue," *Diagnostic Pathology*, vol. 5, no. 1, article 38, 2010.

[18] J. Li, N. Zhang, L.-B. Song et al., "Astrocyte elevated gene-1 is a novel prognostic marker for breast cancer progression and overall patient survival," *Clinical Cancer Research*, vol. 14, no. 11, pp. 3319–3326, 2008.

[19] B. K. Yoo, L. Emdad, Z.-Z. Su et al., "Astrocyte elevated gene-1 regulates hepatocellular carcinoma development and progression," *The Journal of Clinical Investigation*, vol. 119, no. 3, pp. 465–477, 2009.

[20] C. Yu, K. Chen, H. Zheng et al., "Overexpression of astrocyte elevated gene-1 (AEG-1) is associated with esophageal squamous cell carcinoma (ESCC) progression and pathogenesis," *Carcinogenesis*, vol. 30, no. 5, pp. 894–901, 2009.

[21] A. Daponte, M. Ioannou, I. Mylonis et al., "Prognostic significance of hypoxia-inducible factor 1 alpha(HIF-1alpha) expression in serous ovarian cancer: An immunohistochemical study," *BMC Cancer*, vol. 8, article no. 335, 2008.

[22] A. Marín-Hernández, J. C. Gallardo-Pérez, S. J. Ralph, S. Rodríguez-Enríquez, and R. Moreno-Sánchez, "HIF-1α modulates energy metabolism in cancer cells by inducing overexpression of specific glycolytic isoforms," *Mini-Reviews in Medicinal Chemistry*, vol. 9, no. 9, pp. 1084–1101, 2009.

Hindawi
BioMed Research International
Volume 2018, Article ID 4028473, 6 pages
https://doi.org/10.1155/2018/4028473

Research Article

The Prediction of Drug-Disease Correlation Based on Gene Expression Data

Hui Cui,[1,2,3] **Menghuan Zhang**,[2,3] **Qingmin Yang**,[3,4] **Xiangyi Li**,[3] **Michael Liebman**,[3,5] **Ying Yu**,[6] **and Lu Xie**[3]

[1]*School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, China*
[2]*Institute for Nutritional Sciences, Shanghai Institutes for Biological Sciences, University of Chinese Academy of Sciences, Shanghai 200031, China*
[3]*Shanghai Center for Bioinformation Technology, Shanghai Academy of Science and Technology, Shanghai 201203, China*
[4]*College of Food Science and Technology, Shanghai Ocean University, No. 999 Hu Cheng Huan Road, Shanghai 201306, China*
[5]*IPQ Analytics, LLC/Strategic Medicine, Philadelphia, PA, USA*
[6]*Department of Pharmacology, School of Basic Medical Sciences, Tianjin Medical University, Tianjin 30007, China*

Correspondence should be addressed to Ying Yu; yuying@sibs.ac.cn and Lu Xie; luxiex2017@outlook.com

Received 11 November 2017; Revised 18 January 2018; Accepted 11 February 2018; Published 25 March 2018

Academic Editor: Jialiang Yang

The explosive growth of high-throughput experimental methods and resulting data yields both opportunity and challenge for selecting the correct drug to treat both a specific patient and their individual disease. Ideally, it would be useful and efficient if computational approaches could be applied to help achieve optimal drug-patient-disease matching but current efforts have met with limited success. Current approaches have primarily utilized the measureable effect of a specific drug on target tissue or cell lines to identify the potential biological effect of such treatment. While these efforts have met with some level of success, there exists much opportunity for improvement. This specifically follows the observation that, for many diseases in light of actual patient response, there is increasing need for treatment with combinations of drugs rather than single drug therapies. Only a few previous studies have yielded computational approaches for predicting the synergy of drug combinations by analyzing high-throughput molecular datasets. However, these computational approaches focused on the characteristics of the drug itself, without fully accounting for disease factors. Here, we propose an algorithm to specifically predict synergistic effects of drug combinations on various diseases, by integrating the data characteristics of disease-related gene expression profiles with drug-treated gene expression profiles. We have demonstrated utility through its application to transcriptome data, including microarray and RNASeq data, and the drug-disease prediction results were validated using existing publications and drug databases. It is also applicable to other quantitative profiling data such as proteomics data. We also provide an interactive web interface to allow our Prediction of Drug-Disease method to be readily applied to user data. While our studies represent a preliminary exploration of this critical problem, we believe that the algorithm can provide the basis for further refinement towards addressing a large clinical need.

1. Introduction

As we know, many diseases are not resolved by treatment with one single drug, for example, most cancers and diabetes. At time of diagnosis and staging, many aberrant genes can be observed, either involving mutation or modification or exhibiting altered levels of expression, yielding perturbations to signaling pathways. This is the reality of complex diseases, which complicates their treatment particularly in the difficulty in identifying potential driver or passenger genes. Therefore, the traditional "one drug-one target" therapeutic approach often shows limited efficacy because of inappropriate targeting, development of adverse events, and potential resistance [1]. As a result, it has become necessary to develop combination drug therapies [2].

Combined drug therapy typically involves administering two or more drugs simultaneously or sequentially. Within the past two decades, combination therapies have been

used successfully in clinical experiments and have attracted tremendous attention as promising treatments for complex disorders, especially those with multifactorial pathogenic mechanisms [3]. For example, the combination treatment of fluticasone and propionate provides better asthma control than increasing the dose of either single drug alone, while simultaneously reducing the frequency of exacerbations [4]. It is noted that an increasing number of combination drugs are being marketed as commercial products with a fixed dosage of each component and with approval of the Food and Drug Administration (FDA) in the past 5 years, especially for those complex diseases such as type II diabetes, HIV infections, and cancer. In the particular area of cancer therapy, the first combination was granted in January 2014 by FDA to treat melanoma with BRAF V600E or V600K mutations [2]. Currently, approximately 50 combination therapies, without fixed component dosage, have been referred by FDA to treat different cancer subtypes.

Pharmacologically, a drug combination may produce synergistic, additive, antagonistic, or even suppressive effect if the combined effect is greater than, equal to, or less than the sum of each individual drug [5]. Synergistic effects are typically the most desirable because of enhanced efficacy, potential for decreasing dosage with equal or increased level of efficacy, or delayed development of drug resistance [6]. Therefore, identification of synergistic agents presents a significant opportunity to better deal with complex diseases, even though it is a highly challenging task [7]. The synergy of drugs can be assayed by testing the inhibition of tumor cell growth by individual drugs and their combinations in vitro, followed by a mathematical formulation by Loewe additivity or Bliss independence [1, 8]. However, it is not practical to test the synergistic effect of all possible combinations of drugs through experiments due to the large number of drugs approved by FDA. The development of computational methods for predicting effects of drug combination can play an essential role in developing systematic screening of combinatorial treatment regiments [9].

Previous studies have proposed a handful of computational approaches to analyze high-throughput molecular datasets for predicting the synergy of drug combinations. Recently, Zhao et al. introduced a model to predict the efficacies of drug combinations by integrating molecular and pharmacological data. But its dependence on the feature pattern, specifically enriched in approved drug combinations, severely limited its potential application [10]. Similarly, Wu et al. proposed a network-analysis-based model that utilized gene expression profiles, following individual treatments, to predict gene expression changes induced by drug combinations, which were then used to estimate the effectiveness of the combinations [7]. Another model, named the enhanced Petri-Net model, provided informative insight into the mechanisms of drug actions, which was established to recognize the synergism of drug combinations [11]. But its requirement of a gene expression profile for every drug pair limited its application.

However, these computational approaches only consider the characteristics of the drug itself, without taking into account an equivalent characterization of the disease. The effectiveness of the drug may be applicable for the specified cell line, but not applicable for the actual disease as it presents in patients. To account for this, here we propose an algorithm to specifically predict synergistic effects of drug combinations on various diseases, by integrating the data characteristics of disease-related gene expression profiles with drug-treated gene expression profiles. We have demonstrated utility through its application to transcriptome data, including microarray and RNASeq data, and the drug-disease prediction results were validated using existing publications and drug databases. It is also applicable to other quantitative profiling data such as proteomics data. We also provide an interactive web interface (https://www.scbit.org/PEDD/) to allow our Prediction of Drug-Disease method to be readily applied to user data.

2. Methods

In this research, we developed a disease-drug prediction algorithm using transcriptome data. We describe both data aggregation and our algorithm in detail, below.

2.1. Data Aggregation. First, gene expression data of drug treated samples and disease-related gene expression dataset are identified and qualified from literature and public domain databases.

2.1.1. Gene Expression Data following Drug Treatment. GSE51068 dataset (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE51068) was downloaded from the GEO database, which contained gene expression data of 282 drug-treated samples. We selected high-throughput expression profiling of OCI-Ly3 cell line treated with 14 different known drugs at 2 different concentrations and profiled at 6, 12, and 24 hours after treatment. For our initial study, profiling after 6-hour treatment was chosen. Summary information about the 14 known drugs was shown in Table S1.

2.1.2. Disease-Related Gene Expression Data. We have developed our method so that it can be applied not only to microarray data, but also to RNAseq data. Thus, two data types were identified and collected.

We established the following requirements for microarray data in this study: the experimental group involves human disease samples; the control group is nondisease samples; and the number of experimental samples is greater than 50. Six microarray datasets (GSE9476, GSE33615, GSE22529, GSE26049, GSE19429, and GSE47552) were selected from the GEO database (https://www.ncbi.nlm.nih.gov/geo/), including 9 blood cell and bone marrow related malignancies and diseases (Table S2).

Additional disease-related gene expression data involves RNAseq data. Here, four cancer types were chosen including breast cancer, liver cancer, lung adenocarcinoma, and lung squamous cell carcinoma. We extracted these cancer-related RNAseq data from UCSC Xena, which is provided by TCGA (https://xenabrowser.net/datapages/?host=https://tcga.xena-hubs.net).

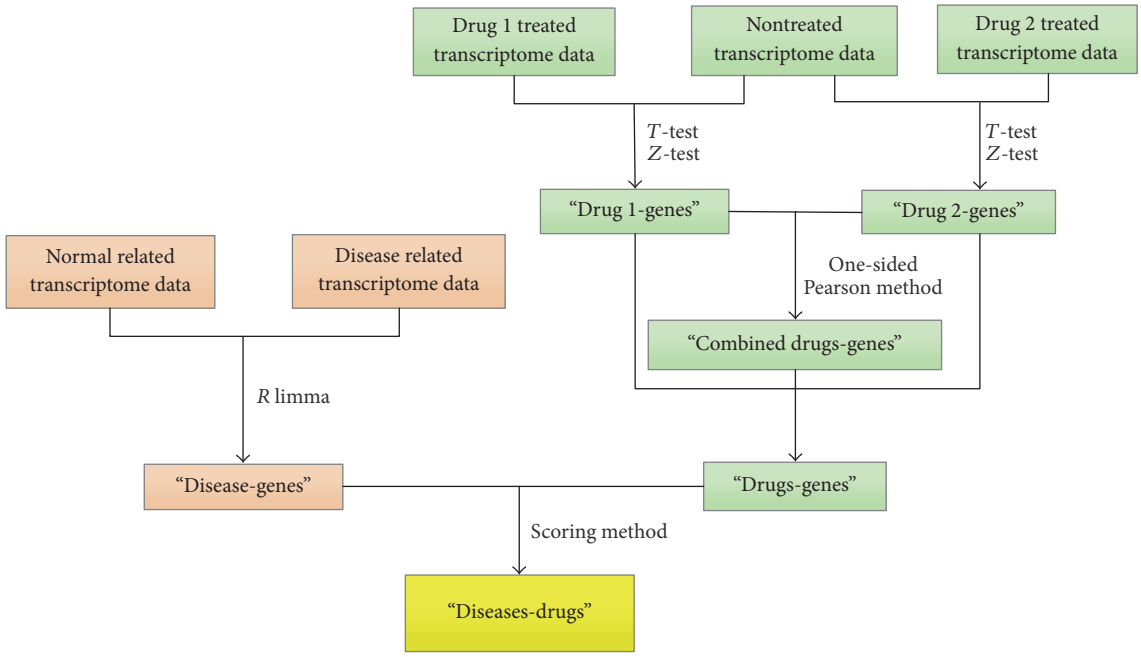


Figure 1: The algorithm flow.

2.2. Algorithm Design and Implementation. Our goal is to predict the effects of drugs on various diseases when used in combination. The detailed algorithm implementation is defined in the steps (Figure 1).

Step 1. Differentially expressed genes (DEGs) were identified within the disease-related gene expression dataset. For microarray data, the "limma" package in *R* was used to identify DEGs, with a Benjamini-Hochberg adjusted p value of 0.01. For RNASeq data, the "limma" package in *R* was also used to identify DEGs, with a Benjamini-Hochberg adjusted p value of 0.05. Additionally, the threshold fold change in gene expression in the experimental group that was selected was at least twice higher or lower than the gene expression in control group for microarray and RNASeq data.

Step 2. DEGs were identified for the 14 drugs. A T test was performed to get the observed test statistics for the genes in the drug-treated group compared to control group. Then, the observed test statistics were converted into z-scores:

$$z_i = \Phi^{-1}\left(P\left(t_i\right)\right), \tag{1}$$

where t_i denotes the observed test statistics for the gene i and $\Phi(\cdot)$ is the cumulative distribution. If the z-score is greater than 1.96, it indicates that the gene expression is upregulated after drug treatment. If the z-score is lower than −1.96, it indicates that the gene expression is downregulated after drug treatment.

Step 3. DEGs were identified for the 91 combination drugs. The 14 drugs will generate 91 unique drug combinations (C_{14}^2). To compute the combined effect of two drugs on each gene, a one-sided Pearson's method was used to combine the z-scores of two drugs:

$$p_i^s = P\left(X_4^2 < -2 \times \sum_{j=1,2} \ln\left(1 - \Phi\left(z_{ij}\right)\right)\right), \tag{2}$$

where z_{ij} $(j = 1, 2)$ denote the z-score of the gene i for any two drugs.

Then, the combined z-score was calculated:

$$z_i' = \Phi^{-1}\left(p_i^s\right). \tag{3}$$

Step 4. DEGs of drug-related and disease-related were matched by evaluating a specific constraint. Here, the p value of the "drug-disease" relationship is calculated using the following formula:

$$p^k = \Phi\left(\frac{\sum_{i=1}^{n} \operatorname{abs}\left(\Phi^{-1}\left(p_i^s\right)\right) I \ \{i \in k\}}{\sqrt{\sum_{i=1}^{k} \operatorname{abs}\left(I\right) \ \{i \in k\}}}\right), \tag{4}$$

where k represents the number of genes that can be matched between the drug and the disease and I is the matching coefficient (Table 1). If the gene is upregulated in the disease and the gene is downregulated after drug treated, I is +1. If the gene is downregulated in the disease and the gene is upregulated after drug treated, I is +1. Otherwise, I is −1.

Table 1: The matching coefficient.

Disease	Drug	
	Up expressed gene	Down expressed gene
Up expressed gene	−1	+1
Down expressed gene	+1	−1

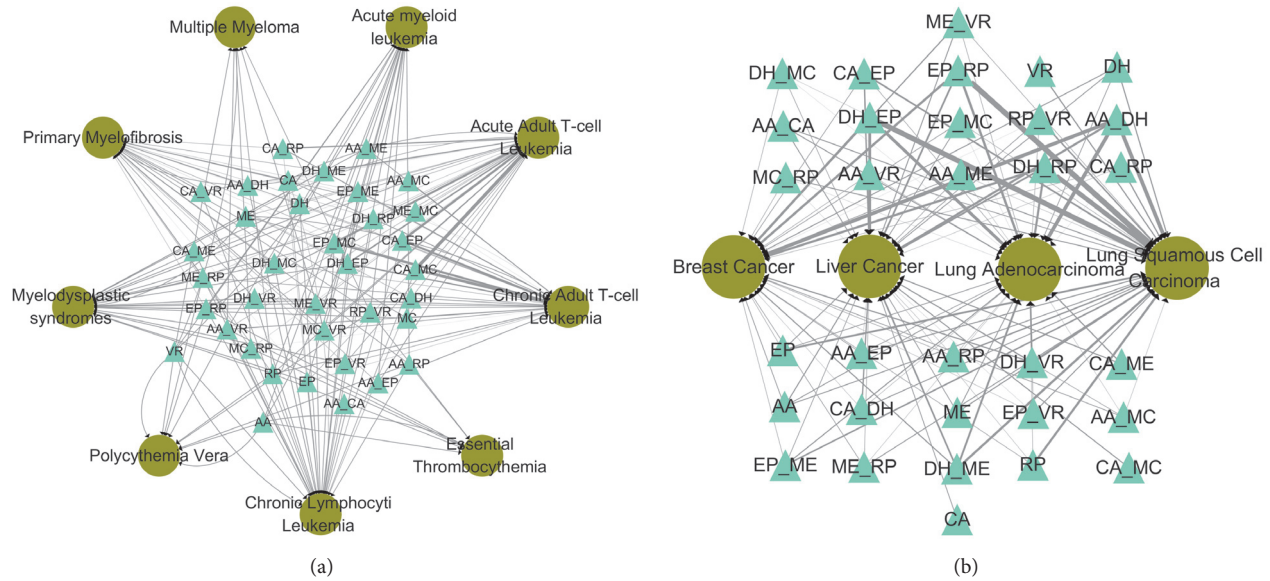


FIGURE 2: The relationship between drug and disease using microarray data (a) and RNASeq data (b). Drugs are represented by triangles. Diseases are represented by circles. The thickness of the linking edge is directly related to the magnitude of the score between drug and disease.

Step 5. An indicator score was calculated, by scoring the matching results, to evaluate the effect of the drug on the disease. The formula is as follows:

$$\text{Score} = \frac{\Phi^{-1}\left(P^k\right) \times k}{N}, \tag{5}$$

where k represents the number of genes that can be matched between the drug and the disease. N is the total number of DEGs in each disease. P is the value calculated in Step 4.

2.3. Literature and Database Validation. For any two drugs (A and B) and any specific disease, three scores can be generated, indicating the relationship between drug A and the disease, between drug B and the disease, and between the A + B drug combination and the disease. Here, we chose the highest score as the most effective. In addition, the score must be greater than 0, suggesting that the drug has an enhanced treatment effect on the disease. If the score of drug combination is higher than that of any single drug, we define the drug combination to be more effective. We chose to exclude those drugs that were not in DrugBank. Finally, results were validated through reviewing both published literature and drug-related databases, including DrugBank (https://www.drugbank.ca/releases/latest) [12], FDA (https://www.fda.Gov/), DCDB (http://www.cls.zju.edu.cn/dcdb/) [13], and the Pubmed (https://www.ncbi.nlm.nih.gov/pubmed).

3. Results

3.1. Relation between Drug and Disease. As a result of our analysis, relationships between drugs and diseases were established and are shown in Figure 2(a) for microarray data. We can see that the most closely related to acute adult T-cell leukemia is the drug combination of camptothecin (CA) and Mitomycin C (MC), followed by the drug combination of camptothecin (CA) and Etoposide (EP) and combination of Etoposide (EP) and Mitomycin C (MC). These drug combinations were also closely related to chronic adult T-cell leukemia, which may be due to their similar pathophysiologic characteristics.

Similarly, relationships between drugs and other cancers are shown in Figure 2(b) for RNASeq data. The drug combination most closely related to breast cancer is that of Aclacinomycin A (AA) and Doxorubicin (DH), followed by the drug combination of Doxorubicin (DH) and Etoposide (EP) and then the combination of Etoposide (EP) and Rapamycin (RP). The most closely related combination to liver cancer involves Doxorubicin (DH) and Etoposide (EP), followed by the drug combination of Aclacinomycin A (AA) and Doxorubicin (DH) and then the combination of Etoposide (EP) and Rapamycin (RP). The drug combination most closely related to lung adenocarcinoma is Aclacinomycin A (AA) and Doxorubicin (DH), followed by the drug combination of Doxorubicin (DH) and Etoposide (EP) and then the combination of Doxorubicin (DH) and Rapamycin (RP). In lung squamous cell carcinoma the most closely related drug combination involves Etoposide (EP) and Rapamycin (RP), followed by the drug combination of Doxorubicin (DH) and Etoposide (EP) and then the combination of Doxorubicin (DH) and Rapamycin (RP).

3.2. Further Validation. As a result of our filtering algorithm (see Methods), a total of 105 relationships between drugs and diseases were identified using microarray data, and a total of 67 relationships were identified using RNASeq data. Then,

results were validated through review of published literature and drug-related databases.

The reviewing identified 36 relationships (microarray) and 41 relationships (RNASeq) in previous studies (Tables S3 and S4). Moreover, there are also 39 synergistic drugs and 18 synergistic drugs identified by previous studies, for microarray and RNASeq data, respectively (Tables S5 and S6).

3.3. Web Interface. We have further implemented the proposed approach as an interactive web tool, named "Predicting the Effect of the Drug on Disease (PEDD)" (https://www.scbit.org/PEDD/). This web tool is intuitive and can be easily applied to similar analyses using user-provided drug-treated gene expression data and disease-related gene expression data, to predict relationships between drugs and diseases. We continue to refine the algorithm and to refine the selection of datasets, for example, both experimental data and disease subtypes, in ongoing studies.

4. Discussion

Due to the complexity of the disease, frequent lack of response to targeted therapies, and the emergence of drug resistance, interest in potential drug combination therapy has increased [14]. Both computational methods and experimental methods have been applied to screen synergistic drugs. An optimal approach would be the potential to use computational screening to broaden the study of potential component drugs for combination therapy and to better direct the application of experimental validation. This approach can lead to more rapid and effective means for screening and identifying candidate drug combinations. Synergistic drug prediction models have been previously studied. For example, Jin et al. built an enhanced Petri-net (EPN) model to predict the synergistic effect of pairwise drug combinations from genome-wide transcriptional expression data, by applying Petri-nets to identify specific drug targeted signaling networks [11]; Sun et al. constructed a model called Ranking-system of Anticancer Synergy (RACS) based on semisupervised learning which was used to rank drug pairs according to their similarity to the labeled samples in a specified multifeature space [15]. However, these computational approaches only considered the characteristics of the drug itself, without taking into account potentially valuable disease observations. The resulting effectiveness of these predictions may be applicable for the cell line, but not readily extendable for disease as it appears in humans. For these reasons, we developed an algorithm to expand on these earlier works and to predict the effects of drugs on various diseases, by integrating gene expression data generated from disease tissues and drug-treated cell lines.

The workflow is as follows. Firstly, up and down genes were calculated with disease-related gene expression data. Secondly, with the gene expression data of drug-treated cell line, we calculated up and down genes for single drug and combination drugs. Next, the disease-related up and down genes were matched with drug-related up and down genes by our matched principle. Moreover, according to the matched result, scores were calculated which represented the effect of drug on various diseases by our scoring method. The implementation of our algorithm as an interactive web tool makes the proposed approach easily accessible to all scientists in general. Researchers can find potential drugs for diseases according to the calculated scores.

In this study, our algorithm can give out the scores of both drug combination and each of the single drug for a disease; thus it is applicable not only to the drug combination prediction, but also to the drug repositioning. Also, according to the score rank, it may be defined that the drug combination is more effective than single drugs if it has the highest score. Besides, this algorithm is not only applicable to transcriptomics data, but also applicable to other quantitative profiling data, such as proteomics data.

The results showed that the effect of combination drugs may be higher than the effect of the individual component drugs in some diseases. For example, the effect of combination of camptothecin and monastrol was predicted to be greater than the effect of camptothecin or monastrol, individually, in acute adult T-cell leukemia and chronic adult T-cell leukemia. In contrast, the effect of combination drugs may be lower than the effect of the individual component drugs in some other diseases. For example, the effect of combination of camptothecin and monastrol was predicted to be reduced in efficacy in multiple myeloma and polycythemia vera. In general, we believe that this analytic approach can contribute to drug research and screening studies and use this preliminary study to show its potential value.

However, in our algorithm, differential genes bear equal weights while the change of some key genes may give larger effect. For example, both gene sequence variations and expression changes are important molecular phenotypes in human disease, especially cancer. They should be assigned differential weights. But, how to determine the key genes and how to assign differential weights for them are very difficult, as we only use the data of gene expression profile in this study. In the future research, more in-depth study of this aspect considering more factors should be carried out. For example, we may use multilevel omics expression data and drug targets to find the key genes and assign differential weights for them. What is more, we also recognize that the disease classes, for example, "breast cancer," that have been used in this study are likely subject to further stratification, for example, DCIS. We are currently studying the application of this approach to such refinements.

And with the rapid development of next-generation sequencing (NGS) technology and the accumulation of histological data [16], there have been many databases that can be used to screen single drugs or synergistic drugs such as FDA and DrugBank [12]. However, a comprehensive database about "drug-cancer relationships" has not been established, which contains both the single drugs and combination drugs related to cancer-related information. We believe such database would be available in future, by collecting the information from current public databases and published literature. The database will provide an important assessment criteria for the "drug-cancer" predictions and provide important reference value for the strategy design of antitumor combination therapy. While our studies represent

a preliminary exploration of this critical direction, we believe that the algorithm can provide the basis for further refinement towards addressing a large clinical need in antitumor combination therapy.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Hui Cui and Menghuan Zhang contributed equally to this work and should be considered co-first authors.

Acknowledgments

This work was supported by National Key Research and Development Program of China [2016YFC0904101], National Natural Science Foundation of China [31570831], National Hi-Tech Program [2015AA020101], and Chinese Human Proteome Projects [CNHPP: 2014DFB30020, 2014DFB30030].

Supplementary Materials

Table S1: drug information. Table S2: microarray data information. Table S3: drug-disease relations identified by previous studies from microarray data. Table S4: drug-disease relations identified by previous studies from RNAseq data. Table S5: synergistic drugs identified by previous studies from microarray data. Table S6: synergistic drugs identified by previous studies from RNAseq data. *(Supplementary Materials)*

References

[1] J. A. Curtin, J. Fridlyand, T. Kageshita et al., "Distinct sets of genetic alterations in melanoma," *The New England Journal of Medicine*, vol. 353, no. 20, pp. 2135–2147, 2005.

[2] Z. Sheng, Y. Sun, Z. Yin, K. Tang, and Z. Cao, "Advances in computational approaches in identifying synergistic drug combinations," *Briefings in Bioinformatics*, 2017.

[3] J. Jia, X. Ma, Z. W. Cao, Y. X. Li, and Y. Z. Chen, "Erratum: Mechanisms of drug combinations: Interaction and network perspectives (Nature Reviews Drug Discovery (2009) vol. 8 (111-128) 10.1038/nrd2683)," *Nature Reviews Drug Discovery*, vol. 8, no. 6, p. 516, 2009.

[4] J. Yang, H. Tang, Y. Li et al., "DIGRE: drug-induced genomic residual effect model for successful prediction of multidrug effects," *CPT: Pharmacometrics & Systems Pharmacology*, vol. 4, no. 2, pp. 91–97, 2015.

[5] P. B. Chapman et al., "Improved survival with vemurafenib in melanoma with BRAF V600E mutation," *The New England Journal of Medicine*, vol. 364, no. 26, pp. 2507-16, 2011.

[6] H. S. Nelson, "Advair: Combination treatment with fluticasone propionate/salmeterol in the treatment of asthma," *The Journal of Allergy and Clinical Immunology*, vol. 107, no. 2, pp. 397–416, 2001.

[7] Z. Wu, X. Zhao, and L. Chen, "A systems biology approach to identify effective cocktail drugs," *BMC Systems Biology*, vol. 4, no. Suppl 2, p. S7, 2010.

[8] M. A. Held, C. G. Langdon, J. T. Platt et al., "Genotype-selective combination therapies for melanoma identified by high-throughput drug screening," *Cancer Discovery*, vol. 3, no. 1, pp. 52–67, 2013.

[9] Q. Xu, Y. Xiong, H. Dai et al., "PDC-SGB: Prediction of effective drug combinations using a stochastic gradient boosting algorithm," *Journal of Theoretical Biology*, vol. 417, pp. 1–7, 2017.

[10] X. Zhao, M. Iskar, G. Zeller, M. Kuhn, V. van Noort, and P. Bork, "Prediction of drug combinations by integrating molecular and pharmacological data," *PLoS Computational Biology*, vol. 7, no. 12, Article ID e1002323, 2011.

[11] G. Jin, H. Zhao, X. Zhou, and S. T. C. Wong, "An enhanced Petri-Net model to predict synergistic effects of pairwise drug combinations from gene microarray data," *Bioinformatics*, vol. 27, no. 13, pp. i310–i316, 2011.

[12] D. S. Wishart, C. Knox, A. C. Guo et al., "DrugBank: a comprehensive resource for in silico drug discovery and exploration," *Nucleic Acids Research*, vol. 34, pp. D668–D672, 2006.

[13] Y. Liu, Q. Wei, G. Yu, W. Gai, Y. Li, and X. Chen, "DCDB 2.0: a major update of the drug combination database," *Database*, vol. 2014, Article ID bau124, 2014.

[14] N. Borisov et al., "A method of gene expression data transfer from cell lines to cancer patients for machine-learning prediction of drug efficiency," *Cell Cycle*, pp. 1–6, 2017.

[15] Y. Sun, Z. Sheng, C. Ma et al., "Combining genomic and network characteristics for extended capability in predicting synergistic drugs for cancer," *Nature Communications*, vol. 6, article 9481, 2015.

[16] J. Reuter, D. V. Spacek, and M. Snyder, "High-throughput sequencing technologies," *Molecular Cell*, vol. 58, no. 4, pp. 586–597, 2015.

Hindawi
BioMed Research International
Volume 2018, Article ID 5747489, 11 pages
https://doi.org/10.1155/2018/5747489

Research Article

SRMDAP: SimRank and Density-Based Clustering Recommender Model for miRNA-Disease Association Prediction

Xiaoying Li[1,2] **Yaping Lin**[1,2] **Changlong Gu**[1] **and Zejun Li**[1,3]

[1]*College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China*
[2]*Key Laboratory of Trusted Computing and Networks, Hunan Province, Changsha 410082, China*
[3]*School of Computer and Information Science, Hunan Institute of Technology, Hengyang 412002, China*

Correspondence should be addressed to Yaping Lin; star@hnu.edu.cn

Received 26 November 2017; Accepted 23 January 2018; Published 21 March 2018

Academic Editor: Tao Huang

Aberrant expression of microRNAs (miRNAs) can be applied for the diagnosis, prognosis, and treatment of human diseases. Identifying the relationship between miRNA and human disease is important to further investigate the pathogenesis of human diseases. However, experimental identification of the associations between diseases and miRNAs is time-consuming and expensive. Computational methods are efficient approaches to determine the potential associations between diseases and miRNAs. This paper presents a new computational method based on the SimRank and density-based clustering recommender model for miRNA-disease associations prediction (SRMDAP). The AUC of 0.8838 based on leave-one-out cross-validation and case studies suggested the excellent performance of the SRMDAP in predicting miRNA-disease associations. SRMDAP could also predict diseases without any related miRNAs and miRNAs without any related diseases.

1. Introduction

MicroRNAs (miRNAs) are small endogenous noncoding RNAs which are approximately 22nt long. Since the discovery of the first two miRNAs lin-4 and let-7, thousands of miRNAs have been identified in eukaryotic cells [1, 2]. A series of studies have shown that miRNAs play an important role in many biological processes, such as cell growth and apoptosis, proliferation, differentiation, and signal transduction [3–6]. Given that miRNAs are involved in the normal function of cells, aberrant miRNA expression has been associated with many types of human diseases, ranging from common diseases to cancers [7–9]. Therefore, the identification of disease-related miRNAs is beneficial in understanding the molecular mechanism of the disease pathogenesis and disease diagnosis and to further promote the level of treatment and prevention.

To date, many biological experimentations have been performed to determine a large number of miRNA-disease associations. Many studies have built databases, such as HMDD [10], miR2Disease [11], dbDEMC [12], miRCancer [13], and PhenomiR [14], to serve as a solid data foundation for predicting miRNA-disease associations. HMDD is a database manually retrieved from the literature [10]. The latest version is HMDD v2.0, which integrates 10,368 miRNA-disease associations of approximately 572 miRNA genes and 378 diseases from 3,511 papers. MiR2Disease documents 1,939 manually curated miRNA-disease associations between 299 human miRNAs and 94 human diseases [11]. The dbDEMC stores differentially expressed miRNAs in human cancers obtained from microarray data [12]. The updated version dbDEMC 2.0 contains 2,224 differentially expressed miRNAs in 36 cancer types [15]. The miRCancer stores miRNA-cancer associations obtained by text mining method [13]. PhenomiR provides information about differentially regulated miRNA expression in diseases and other biological processes [14].

However, using experimental methods to identify the disease-related miRNAs is time-consuming and costly. Based on existing data, computational methods have been developed as a valuable supplement to the experimental methods to save experimental time and cost. Computational methods can calculate and rank the similarity scores of all miRNAs for a given disease. Top-ranked miRNAs are treated as the most promising candidate disease miRNAs for further

experimental studies. Similarity calculation is the key issue in computational methods [16]. According to the calculation of similarity score, most computational methods are divided into two categories [17, 18], namely, network-based methods [19–28] and machine-learning-based methods [24, 29–34]. Network-based methods predict miRNA-disease associations by considering the hypothesis that miRNAs with similar functions usually tend to be associated with phenotypically similar diseases [10]. Jiang et al. [19] constructed a human phenome-miRNAome functional association miRNA network using the hypergeometric distribution scoring system to select the candidate disease miRNAs. However, high final prediction accuracy may not be obtained if only the local information of each miRNA is issued and the study is strongly dependent on the predicted miRNA-target interactions. Chen et al. [21] adopted global network similarity measures and developed RWRMDA to infer the associations between diseases and miRNAs by implementing random walk on the miRNA-miRNA function similarity network. Based on the weighted k most similar neighbors, Xuan et al. [22] proposed HDMP to infer disease-related miRNAs. HDMP evaluates miRNA function similarity by incorporating the information content of disease terms, disease phenotype similarity, and weight information of the miRNA family or cluster. However, RWRMDA and HDMP cannot be useful for predicting disease without any related miRNAs. Based on social network analysis, Zou et al. [24] proposed KATZ method to compute the similarity score based on walks of different lengths between the miRNA and disease nodes. However, KATZ has relatively poor capability of sparing known associations. Gu et al. [25] calculated miRNA similarity and disease similarity of known miRNA-disease associations through the Jaccard similarity measure. They incorporated miRNA similarity of known miRNA-disease associations, miRNA functional similarity, and miRNA family information to construct miRNA similarity network and incorporated disease similarity of known miRNA-disease associations to construct disease similarity network. Then, they applied network consistency projection method to predict the disease-related miRNAs.

Machine-learning-based methods extract features from data to initially obtain effective features of miRNAs and diseases and then utilize machine learning models to predict miRNA-disease associations. Jiang et al. [29] showed a support vector machine (SVM) classifier method by integrating the feature vectors of miRNA-target and phenotype similarity. Xu et al. [31] introduced an approach based on the miRNA-target-dysregulated network to prioritize novel disease miRNAs. This method also constructs a support vector machine classifier based on the features and changes in miRNA expression. However, these two computational methods are mainly limited by the difficulty or impossibility of obtaining negative training samples, and this drawback would largely influence the predictive accuracy. To solve this problem, Chen and Yan [30] developed a semisupervised method of regularized least squares for miRNA-disease association (RLSMDA). RLSMDA integrates known disease-miRNA associations, disease similarity dataset, and miRNA functional similarity network to infer potential disease-related miRNAs. The main drawback of RLSMDA is the intricate adjustment of parameters. Xiao et al. [35] used graph-regularized nonnegative matrix factorization framework to predict potential miRNA-disease associations using weighted k nearest neighbor profiles to incorporate miRNA similarity and disease matrices. Chen et al. [34] presented a computational method DRMDA based on stacked autoencoder, greedy layer-wise unsupervised pretraining algorithm and SVM, and this method was implemented to predict potential miRNA-disease associations. However, DRMDA results are not highly accurate, because of the difficulty in obtaining negative samples and optimizing the complex parameters.

Similarity calculation mainly considers miRNA-miRNA similarity measurement. Several computational methods use the known miRNA-disease associations in calculating miRNA-miRNA similarity [19–26, 29, 30]. In these methods, miRNA-miRNA similarity measurement is completed by disease-disease measurement and known experimental miRNA-disease associations. However, these methods are restricted by the possible overestimation of the predictive accuracy. This drawback may be due to the fact that cross-validation experiments are not correctly performed, and the miRNA-miRNA similarity depends heavily on the known miRNA-disease associations. These methods fail to remove known information of the tested element for similarity calculation at each round of cross-validation. Other limitations include the inability to predict isolated miRNA and lack of disease semantic similarity [36]. An isolated miRNA signifies that a miRNA has no associated disease; that is, no relationship exists between this isolated miRNA and diseases. Thus, miRNA-disease associations cannot be used to calculate miRNA similarity of an isolated miRNA. Instead of using experimentally verified miRNA-disease associations, other computational methods calculate miRNA similarity using the interaction of miRNAs with other biomolecules [31, 36–38]. For example, Liu et al. [36] calculated miRNA similarity using the miRNA-target gene and miRNA-long noncoding RNA associations. However, the performances of these methods are deficient.

Based on the assumption that miRNAs with similar functions are normally associated with phenotypically similar diseases and vice versa, we solved the aforementioned limitations by establishing a novel computational method based on SimRank [39] and density-based clustering [40] recommender model for miRNA-disease association prediction (SRMDAP). The SRMDAP constructs miRNA similarity subnetwork using SimRank to calculate network topological similarity between miRNAs based on miRNA-message RNA (mRNA) interaction network. The disease similarity subnetwork is similar to miRNA similarity subnetwork and is based on the disease-gene network. Then, the SRMDAP uses the density-based clustering recommender model to integrate miRNA similarity subnetwork, disease similarity subnetwork, and experimentally verified miRNA-disease associations to predict potential associations between miRNAs and diseases. In this work, leave-one-out cross-validation experiment and case studies about two important cancers, namely, kidney and colorectal neoplasms, have

indicated the excellent predictive performance of SRMDAP. The SRMDAP can also predict isolated diseases and isolated miRNAs.

2. Methods

2.1. Data. Three datasets were used in our approach. Experimentally verified miRNA-mRNA interactions were downloaded from the miRTarBase database to construct the miRNA similarity network [41] (http://mirtarbase.mbc.nctu.edu.tw/, Release 6.0: Sept-15-2015). Meanwhile, experimentally verified disease-related mRNAs were downloaded from the DisGeNET database [42] (http://www.disgenet.org/web/DisGeNET/menu/home, DisGeNET 4.0: October 2016) to construct a disease similarity network. Experimentally verified miRNA-disease network was downloaded from the HMDD v2.0 database [43] (http://www.cuilab.cn/hmdd, Jun-14-2014 Version).

2.2. Data Processing

2.2.1. MiRNA-Disease Association Network. The disease names of the DisGeNET and HMDD databases were mapped to the MeSH description (https://www.ncbi.nlm.nih.gov/mesh). Diseases in the HMDD database not found in the DisGeNET database and repeated associations were removed. Then, we obtained 5,048 known miRNA-disease associations, including 475 miRNAs and 334 diseases, as the benchmark dataset. Formally, we denoted the miRNA set as $\mathbf{M} = \{\mathbf{m_1}, \mathbf{m_2}, \ldots, \mathbf{m_{nm}}\}$ and the disease set as $\mathbf{D} = \{\mathbf{d_1}, \mathbf{d_2}, \ldots, \mathbf{d_{nd}}\}$. The variables **nm** and **nd** denote the number of miRNAs and diseases, respectively. Matrix **AS** represents the adjacency matrix of miRNA-disease associations. $\mathbf{AS(i, j)} = \mathbf{1}$ denotes miRNA **i** associated with disease **j**; otherwise, $\mathbf{AS(i, j)} = \mathbf{0}$.

2.2.2. MiRNA Similarity Network. SimRank [39] was employed to calculate the disease and miRNA similarities based on miRNA-mRNA interaction network and disease-related mRNA associations. SimRank is a model to measure the degree of similarity between any two objects on the basis of the information of the topology graph, which has been successfully applied to web page ranking [44], recommender systems [45], outlier detection [46], network graph clustering [47], and approximate query processing [48], among others. The SimRank model defines the similarity of two nodes based on a recursive thinking. When other nodes pointing to the two nodes are similar, then the two nodes are similar. SimRank defines the similarity of two nodes as follows:

$$\mathbf{s}(\mathbf{a}, \mathbf{b}) = \begin{cases} \mathbf{1} & \mathbf{a} = \mathbf{b} \\ \dfrac{\mathbf{C}}{|\mathbf{I}(\mathbf{a})| \cdot |\mathbf{I}(\mathbf{b})|} \sum_{\mathbf{j} \in \mathbf{I(b)}} \sum_{\mathbf{i} \in \mathbf{I(a)}} \mathbf{s}(\mathbf{i}, \mathbf{j}) & \mathbf{a} \neq \mathbf{b} \\ \mathbf{0} & \mathbf{I}(\mathbf{a}) = \emptyset \text{ or } \mathbf{I}(\mathbf{b}) = \emptyset, \end{cases} \tag{1}$$

where $\mathbf{s}(\mathbf{a}, \mathbf{b})$ is the similarity between nodes **a** and **b** and $\mathbf{C} \in [\mathbf{0}, \mathbf{1}]$ is a decay factor. $\mathbf{I}(\mathbf{a})$ denotes all node sets that point to node **a**, and $|\mathbf{I}(\mathbf{a})|$ is the number of elements of $\mathbf{I}(\mathbf{a})$.

The adjacency matrix of the miRNA-mRNA interaction bipartite network is represented as **A**, where $\mathbf{A(i, j)}$ in row **i** and column **j** is 1 if miRNA **i** is associated with mRNA **j**, and 0 otherwise. The matrix **A** is normalized by column to determine the matrix $\mathbf{W_1}$, and the similarity matrix can be calculated as follows:

$$\mathbf{SM} = \mathbf{C_1} \cdot \left(\mathbf{W_1}^{\mathbf{T}} \cdot \mathbf{SM} \cdot \mathbf{W_1}\right) + \left(\mathbf{1} - \mathbf{C_1}\right) \cdot \mathbf{I}, \tag{2}$$

where **SM** is the miRNA similarity matrix and $\mathbf{SM(i, j)}$ is the similarity between miRNAs **i** and **j**. $\mathbf{W_1}^{\mathbf{T}}$ is the transpose matrix of $\mathbf{W_1}$, $\mathbf{C_1}$ is a decay factor, and **I** is the unit matrix.

2.2.3. Disease Similarity Network. We can obtain the similarity matrix of diseases using the same process in determining the miRNA similarity network. The adjacency matrix of the disease-gene network is represented as **B**, where $\mathbf{B(i, j)}$ in row **i** and column **j** is 1 if the disease **i** is associated with gene **j**, and 0 otherwise. Matrix **B** is normalized by column to obtain the matrix $\mathbf{W_2}$, and the similarity matrix can be calculated as follows:

$$\mathbf{SD} = \mathbf{C_2} \cdot \left(\mathbf{W_2}^{\mathbf{T}} \cdot \mathbf{SD} \cdot \mathbf{W_2}\right) + \left(\mathbf{1} - \mathbf{C_2}\right) \cdot \mathbf{I}, \tag{3}$$

where **SD** is the disease similarity matrix and $\mathbf{SD(i, j)}$ is the similarity between diseases **i** and **j**. $\mathbf{W_2}^{\mathbf{T}}$ is the transpose matrix of $\mathbf{W_2}$, $\mathbf{C_2}$ is a decay factor, and **I** is the unit matrix. A simple example of constructing miRNA and disease similarity is provided in Figure 1.

2.3. Prediction Method. In this work, a density-based clustering recommendation model is developed based on the miRNA and disease similarity network to predict potential miRNA-disease associations. The flowchart of SRMDAP is shown in Figure 2.

For example, the calculation for predicting the association of miRNA i and disease j is as follows. First, given the assumption that miRNAs with similar functions are normally associated with phenotypically similar diseases and vice versa [10, 49], the closer the neighbors of miRNA i are to disease j, the closer miRNA i will be to disease j in the miRNA similarity network. Using miRNA i as cluster center and greedy method, we added the most similar neighbor nodes to form new clusters, until the cluster density no longer increased. The cluster density of cluster V is defined as follows:

$$d(V) = \frac{w^{\text{in}}}{w^{\text{in}} + w^{\text{out}} + \alpha \cdot |V|}, \tag{4}$$

where w^{in} and w^{out} denote the sum of the weights of inner and external sides of cluster V, respectively [50]. Item $\alpha \cdot |V|$ is a penalty item, and $|V|$ is the number of members of cluster V. In our experiments, we set $\alpha = 2$. Then, using $V_m(i) = [m_1, m_2, \ldots, m_n]$, which denotes the closest neighbors of miRNA i, the predictive score between miRNA i and disease j is calculated as follows:

$$RS1(i, j) = \frac{\sum_{k \in V_m(i)} SM(i, k) \cdot AS(k, j)}{|V_m(i)|}, \tag{5}$$

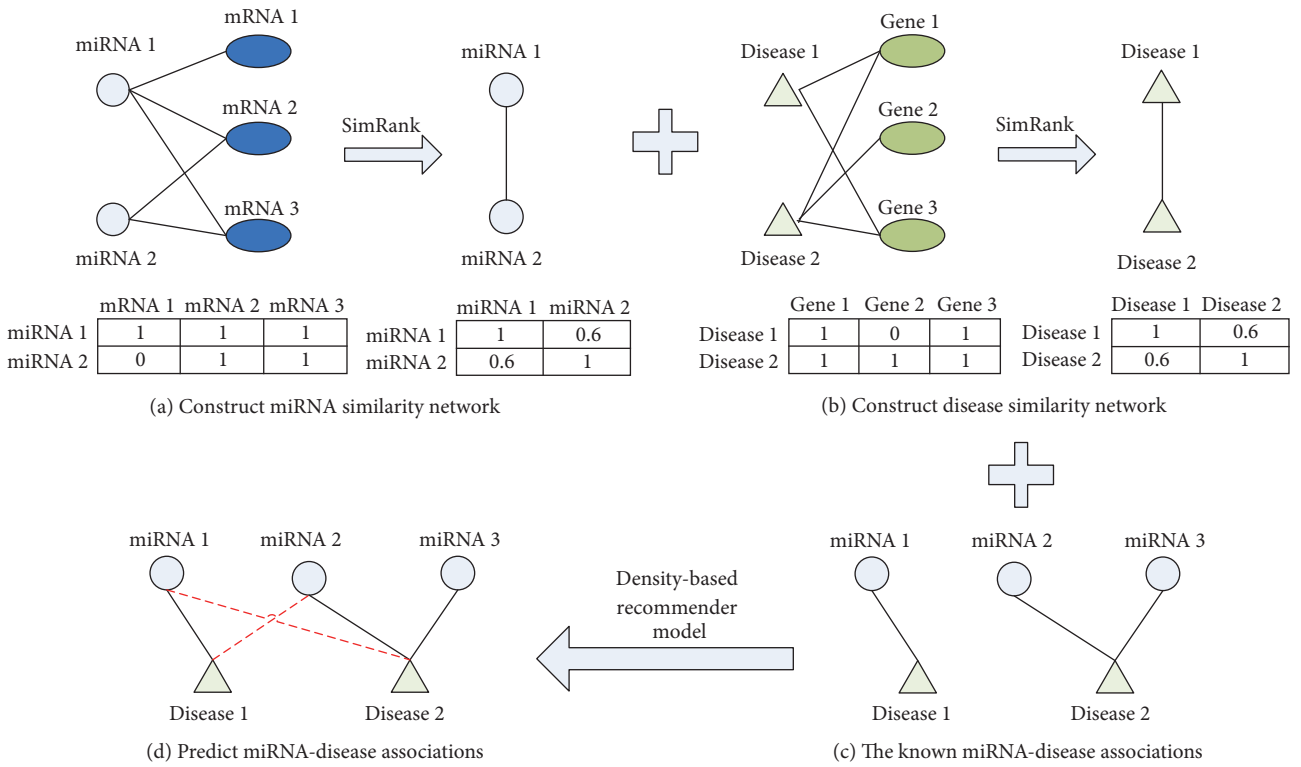


FIGURE 1: Illustration of the process of constructing miRNA and disease similarity network and predicting miRNA-disease associations. (a) A simple example of constructing similarity of miRNAs 1 and 2 is shown in (a). (b) A simple example of constructing similarity of diseases 1 and 2 is shown in (b). (c) The known miRNA-disease associations. (d) Predicting miRNA-disease associations through density-based recommender model by integrating miRNA similarity network, disease similarity network, and the known miRNA-disease associations.

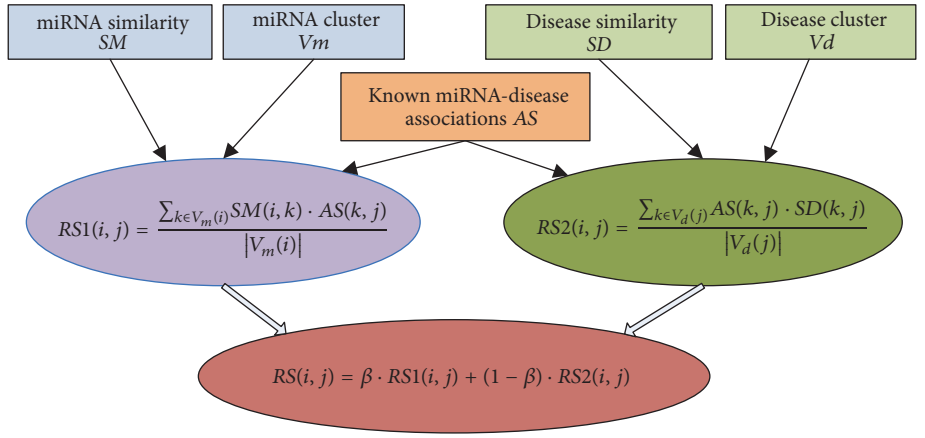


FIGURE 2: The flowchart of SRMDAP.

where $RS1(i, j)$ is the predictive score between miRNA i and disease j calculated by the neighbors of miRNA i; and $SM(i, k)$ is the similarity of miRNA i and miRNA k; and $AS(k, j)$ is the association between miRNA k and disease j. Equation (5) calculates the predictive score based on the nearest neighbors of miRNA i and the associations between the neighbors and disease j.

Second, in the same way, based on the assumption that diseases with similar functions often have similar semantic descriptions and vice versa [20], the closer the neighbors of disease j are to miRNA i, the closer the disease j will be to miRNA i in the disease similarity network; the predictive score between miRNA i and disease j is calculated as follows:

$$RS2(i,j) = \frac{\sum_{k \in V_d(j)} AS(k,j) \cdot SD(k,j)}{|V_d(j)|}, \quad (6)$$

where $V_d(j)$ is the closest neighbor to disease j.

Table 1: Global characteristic of the known miRNA-disease association network.

Number of diseases	Number of miRNAs	Number of miRNA-disease association	Avg. degree of diseases	Avg. degree of miRNAs	Max degree of diseases	Min degree of disease	Max degree of miRNAs	Min degree of miRNAs
334	475	5048	15.11	10.63	208	1	112	1

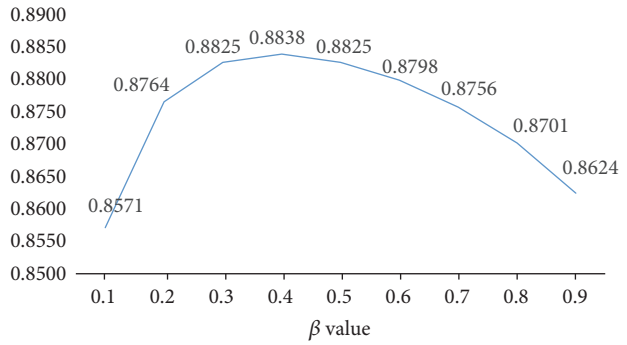


Figure 3: Average AUCs affected by β value. When β is 0.4, average AUC is 0.8838 and SRMDAP achieves the best performance.

Finally, the final predictive score between miRNA i and disease j is calculated by integrating $RS1(i, j)$ and $RS2(i, j)$ as follows:

$$RS(i, j) = \beta \cdot RS1(i, j) + (1 - \beta) \cdot RS2(i, j), \quad (7)$$

where $\beta \in [0, 1]$ is an integration parameter to balance the contributions from miRNA and disease similarities. $RS(i, j)$ in row i and column j is the prediction value of miRNA i to disease j.

When the predictive score between isolated disease j and miRNA i is calculated, all associations of isolated disease j are ignored, and the contribution of the neighbors of miRNA i to the predictor is zero. Thus, $RS1(i, j)$ equals 0. The final predictive score between isolated disease j and miRNA i is $RS2(i, j)$, which is the predictive score between the similarity neighbors of disease j and miRNA i. Therefore, SRMDAP can predict associated miRNAs for an isolated disease. Similarly, when the predictive score between new miRNA and disease is calculated, $RS1(i, j)$ is the predictive score between the similarity neighbors of miRNA i and disease j, and only $RS1(i, j)$ is used as the predictive score between the new miRNA and related diseases.

To explore for a suitable β value, we tested different β values from 0.1 to 0.9 and calculated the average area under the curve (AUC) in the framework of leave-one-out cross-validation. The results showed that SRMDAP achieved the highest average AUCs when β was 0.4 (Figure 3).

3. Results

3.1. Characteristics of the miRNA-Disease Association Network. In our study, 5,048 known miRNA-disease associations consisting of 475 miRNAs and 334 diseases were included. To comprehensively illustrate the known miRNA-disease association network, we demonstrated the characteristics of known miRNA-disease association network in Table 1. The degree of a disease (or miRNA) represented the neighboring miRNAs (or disease) related to it. The average degrees of the disease and miRNAs were 15.11 and 10.63, respectively. The degree of distribution of diseases and miRNAs of the known miRNA-disease association network (Figure 4) revealed a power-law distribution. Most of the miRNAs and diseases presented a degree of 1. Hepatocellular carcinoma showed that the maximum degree, that is, 208 miRNAs, was related to this malignancy. Meanwhile hsa-mir-21 showed the maximum degree, with 112 diseases related to this miRNA.

3.2. Performance Evaluation of SRMDAP. We implemented the leave-one-out cross-validation (LOOCV) on the known miRNA-disease associations to evaluate the predictive performance of the SRMDAP. For a given disease d, each known association between miRNA and disease d was ignored in turn as a test sample, and other known associations between miRNAs and disease d were considered as a training set. The remaining miRNAs without evidence to show their relation to disease d composed the candidate miRNA set. We calculated the relevance score of these candidate miRNAs with disease d and ranked them by their scores. If the rank exceeded a given threshold, then the SRMDAP model successfully predicted this miRNA-disease association. The threshold was varied to draw the receiver operating characteristic (ROC) curve, and the score of the AUC was calculated to demonstrate the predictive performance. The ROC plots the relationship between the true positive rate (TPR, sensitivity) and the false positive rate (FPR, 1 − specificity) at different thresholds. Sensitivity represents the percentage of test miRNA-disease associations with ranking above a given threshold. Meanwhile, specificity represents the percentage of miRNA-disease associations below the threshold.

The TPR and FPR were calculated as follows:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \qquad \text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}}, \quad (8)$$

where TP, FP, TN, and FN indicate true positive, false positive, true negative, and false negative, respectively. Given a threshold, TP and FP are the number of known and unknown associations above the threshold, respectively. TN and FP are the number of unknown and known associations below the threshold, respectively. The AUC value of 1 indicates perfect performance of the prediction method. Moreover, an

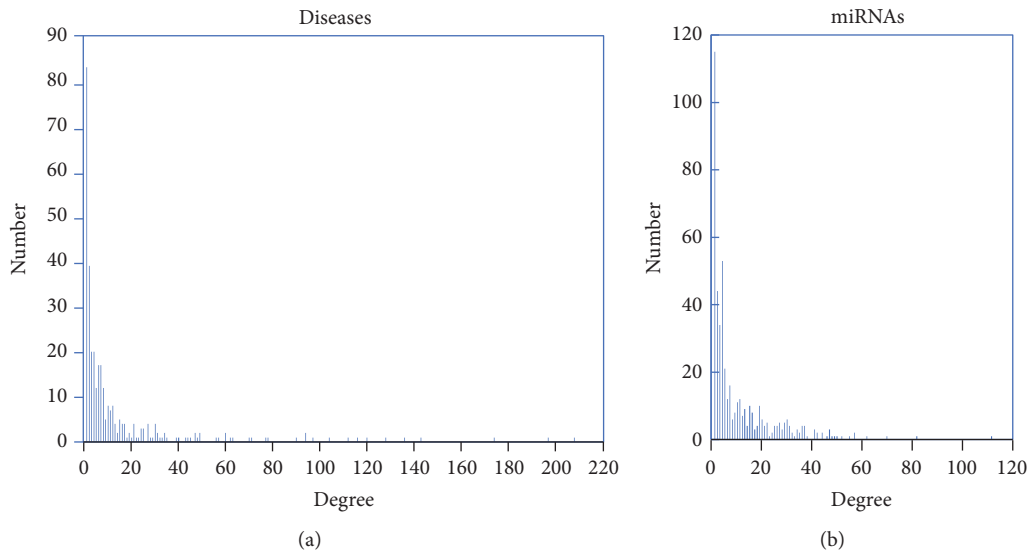


Figure 4: Disease degree distribution and miRNAs degree distribution in the known miRNA-disease association network. (a) shows the bar diagram of disease degree. (b) shows the bar diagram of miRNAs degree.

AUC value of 0.5 implies the random performance of the prediction method.

To our knowledge, RLSMDA [30], KATZ [24], and Liu et al.'s method [36] are three the-state-of-the-art computation methods that predict miRNA-disease associations. In our work, we compared SRMDAP with these methods and implemented a LOOCV for the three methods. The SRMDAP achieved the highest AUC of 0.8838 when $\beta = 0.4$. When optimal parameters were selected as described by the authors, AUC values corresponding to RLSMDA, KATZ, and Liu's method were 0.8584, 0.8522, and 0.7983, respectively. Comparative results of overall ROC curves and AUCs of all methods are shown in Figure 5.

To obtain a reliable judgment, we tested 18 human diseases associated with at least 70 miRNAs, because diseases related to a few miRNAs were not sufficient to evaluate the performance of the prediction methods. Table 2 shows that the SRMDAP achieved the highest AUC of 0.8874 with lung neoplasms and lowest AUC of 0.7367 with renal cell carcinoma. The average AUC value for the 18 diseases was 0.8056. The average AUC values for the 18 diseases obtained from RLAMDA, KATA, and Liu's method were 0.6671, 0.6901, and 0.5178, respectively. The average AUC achieved by SRMDAP was 14%, 12%, and 29% higher than those of the other three methods, respectively. The AUC values of the SRMDAP for the 18 diseases were all higher than those of RLSMDA, KATZ, and Liu's method. These facts indicated that the prediction performance of SRMDAP was superior to RLSMDA, KATZ, and Liu's method.

3.3. Case Studies. To further evaluate the SRMDAP's ability to discover potential miRNA-disease associations, we selected two important diseases (kidney neoplasms and colorectal neoplasms) as case studies. We analyzed the top 50 candidates in detail. Prediction results were supported by dbDEMC [15] database and literature.

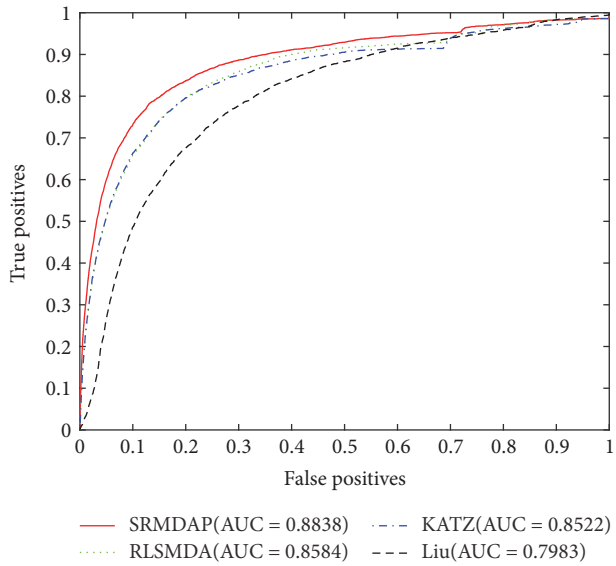


Figure 5: Method comparison: comparison between SRMDAP, RLSMDA, KATZ, and Liu's method in terms of ROC curve and AUC.

Kidney neoplasm, which forms in tissues of the kidneys, is one of the top 10 cancer killers. This malignancy is still difficult to diagnose and treat. Based on 2010–2014 cases and deaths, the annual number of new cases of kidney and renal pelvis cancer was 15.6 per 100,000 persons. The five-year survival rate in the United State is 74.1% [51]. MiRNAs showing altered expression in the kidney are promising biomarkers for diagnosis. For example, miR-141 and miR-200b are underexpressed in renal cell carcinoma (a kidney neoplasm type) from normal kidney and oncocytoma in tissue samples. The miRNA expression profiles of miR-141 or miR-200b might provide an ancillary tool for the

Table 2: Prediction result of SRMDAP and other methods for LOOCV.

Disease names	Number of related_miRNAs	AUC			
		SRMDAP	RLSMDA	KATZ	Liu's method
Carcinoma, hepatocellular	208	**0.7639**	0.6909	0.6881	0.4807
Breast neoplasms	197	**0.7776**	0.6814	0.6779	0.4147
Stomach neoplasms	174	**0.7591**	0.6635	0.6791	0.5498
Colorectal neoplasms	143	**0.7929**	0.6647	0.6895	0.4699
Melanoma	136	**0.7958**	0.6584	0.6673	0.4804
Lung neoplasms	128	**0.8874**	0.7198	0.7675	0.5243
Heart failure	120	**0.7538**	0.6608	0.6622	0.5040
Prostatic neoplasms	116	**0.8076**	0.6704	0.7054	0.5440
Ovarian neoplasms	112	**0.8732**	0.7194	0.7705	0.5382
Carcinoma, renal cell	104	**0.7367**	0.5815	0.6126	0.4932
Pancreatic neoplasms	97	**0.8687**	0.6829	0.7288	0.5355
Carcinoma, non-small-cell lung	94	**0.8322**	0.6873	0.6981	0.5470
Glioblastoma	94	**0.7686**	0.6421	0.6522	0.5644
Urinary bladder neoplasms	90	**0.7935**	0.6231	0.6635	0.5475
Carcinoma, squamous cell	78	**0.8637**	0.7179	0.7200	0.5398
Colonic neoplasms	77	**0.8271**	0.6582	0.6859	0.5490
Glioma	71	**0.8212**	0.6727	0.7146	0.5591
Esophageal neoplasms	70	**0.7789**	0.6126	0.6383	0.4781

correct discrimination of kidney neoplasms [52]. Candidate miRNAs were ranked based on the SRMDAP. The top 50 potential miRNAs associated with kidney neoplasms and evidence for the associations with kidney are listed in Table 3. Among the top 50 predicted candidates, 49 miRNA have been confirmed by dbDEMC, and only hsa-mir-7 is not confirmed by dbDEMC. However, downregulation of miR-7 with synthesized inhibitor inhibited cell migration in vitro, suppressed cell proliferation, and induced renal cancer cell apoptosis. Thus, miR-7 could be characterized as an oncogene in renal cell carcinoma [53].

Colorectal neoplasm is the third most common cancer and the fourth most common cancer-related cause of death worldwide, with more than 1.2 million new cases and 600,000 deaths annually [54]. MiRNAs can be used as useful biomarkers for colorectal cancer diagnosis, prognosis, and prediction of treatment response because of their several unique characteristics [55]. For example, serum miR-21, miR-29a, and miR-125b levels could discriminate early colorectal neoplasms patients from healthy controls [56]. The top 50 potential miRNAs associated with colorectal neoplasms and evidence for associations with kidney are listed in Table 4. Among the top 50 predicted candidates, 49 miRNAs were confirmed by dbDEMC. Only 1 miRNA (hsa-mir-663a) was not confirmed in the dbDEMC.

3.4. Prediction of Isolated Diseases and Isolated miRNAs. An isolated disease signifies a disease without any known related miRNAs or newly discovered disease. When we tested the capability of SRMDAP to predict isolated diseases, we removed all known verified miRNAs, which have been shown to be related to the predicted disease. This operation was performed to confirm that we only used the similarity information of other miRNAs-related diseases to predict candidate miRNAs associated with the given disease. Then, these candidate miRNAs were ranked according to their scores. The average AUC of SRMDAP to predict isolated disease was 0.7990. For colorectal neoplasms, we removed 143 known miRNA related to colorectal neoplasms and ranked candidate miRNAs based on the predictive result of SRMDAP. Among the top 50 predicted candidates, 49 miRNAs have been confirmed by dbDEMC. The potential candidate hsa-mir-494 is supported by the literature [PMID: 25270723]. However, hsa-mir-494 is an independent prognostic marker for colorectal neoplasm patients, and this miRNA promotes cell migration and invasion in colorectal neoplasms by directly targeting PTEN [57]. The predicted results of colorectal neoplasms are listed in Table 5.

As previously stated, an isolated miRNA is a miRNA without any known related disease, such as newly discovered miRNAs. The known verified disease-miRNA associations related to predictive miRNAs were removed to demonstrate the ability of SRMDAP to predict miRNAs without any known related disease. This procedure ensures the use of only known disease-miRNA associations and similarity information of other miRNAs to predict candidate disease. Then, these candidate diseases were ranked according to their scores. The average AUC of the SRMDAP to predict isolated miRNAs was 0.8464. The predicted results of hsa-mir-106b are listed in Table 6. For hsa-mir-106b, we removed 31 related diseases associations and ranked candidate diseases based on the predictive result of the SRMDAP. Among the top 10 predicted candidates, all diseases have been confirmed by dbDEMC, miR2Disease, or HMDD. These results demonstrate that the SRMDAP may be recommended to predict isolated diseases and miRNAs.

Table 3: The top 50 potential kidney neoplasms-related miRNAs predicted by SRMDAP and the confirmation of these associations. Forty-nine of the top 50 kidney neoplasms-related miRNAs have been confirmed by dbDEMC. Hsa-mir-7 ranked 48th has been confirmed by the literature (PMID: 23793934).

Rank	miRNA	Evidence
(1)	hsa-mir-155	dbDEMC
(2)	hsa-mir-146a	dbDEMC
(3)	hsa-mir-17	dbDEMC
(4)	hsa-mir-125b	dbDEMC
(5)	hsa-mir-20a	dbDEMC
(6)	hsa-mir-34a	dbDEMC
(7)	hsa-mir-145	dbDEMC
(8)	hsa-mir-92a	dbDEMC
(9)	hsa-mir-16	dbDEMC
(10)	hsa-mir-126	dbDEMC
(11)	hsa-mir-18a	dbDEMC
(12)	hsa-mir-221	dbDEMC
(13)	hsa-mir-19b	dbDEMC
(14)	hsa-mir-29a	dbDEMC
(15)	hsa-mir-1	dbDEMC
(16)	hsa-mir-29b	dbDEMC
(17)	hsa-let-7a	dbDEMC
(18)	hsa-mir-19a	dbDEMC
(19)	hsa-mir-143	dbDEMC
(20)	hsa-mir-223	dbDEMC
(21)	hsa-mir-200b	dbDEMC
(22)	hsa-mir-29c	dbDEMC
(23)	hsa-mir-31	dbDEMC
(24)	hsa-let-7b	dbDEMC
(25)	hsa-mir-222	dbDEMC
(26)	hsa-mir-181a	dbDEMC
(27)	hsa-mir-210	dbDEMC
(28)	hsa-mir-199a	dbDEMC
(29)	hsa-mir-200a	dbDEMC
(30)	hsa-mir-133a	dbDEMC
(31)	hsa-mir-150	dbDEMC
(32)	hsa-mir-34c	dbDEMC
(33)	hsa-mir-146b	dbDEMC
(34)	hsa-let-7c	dbDEMC
(35)	hsa-mir-142	dbDEMC
(36)	hsa-mir-181b	dbDEMC
(37)	hsa-mir-124	dbDEMC
(38)	hsa-mir-9	dbDEMC
(39)	hsa-mir-106b	dbDEMC
(40)	hsa-let-7e	dbDEMC
(41)	hsa-mir-133b	dbDEMC
(42)	hsa-mir-196a	dbDEMC
(43)	hsa-mir-182	dbDEMC
(44)	hsa-let-7d	dbDEMC
(45)	hsa-mir-30a	dbDEMC
(46)	hsa-mir-148a	dbDEMC
(47)	hsa-mir-195	dbDEMC
(48)	hsa-mir-7	**PMID: 23793934**
(49)	hsa-mir-34b	dbDEMC
(50)	hsa-mir-24	dbDEMC

Table 4: The top 50 potential colorectal neoplasms-related miRNAs predicted by SRMDAP and the confirmation of these associations. Forty-nine of the 50 colorectal neoplasms-related miRNAs have been confirmed by dbDEMC. Only 1 miRNA (hsa-mir-663a is ranked 30th) is unconfirmed.

Rank	miRNA	Evidence
(1)	hsa-mir-650	dbDEMC
(2)	hsa-mir-15a	dbDEMC
(3)	hsa-mir-223	dbDEMC
(4)	hsa-mir-29b	dbDEMC
(5)	hsa-mir-518b	dbDEMC
(6)	hsa-mir-192	dbDEMC
(7)	hsa-mir-488	dbDEMC
(8)	hsa-mir-29c	dbDEMC
(9)	hsa-mir-521	dbDEMC
(10)	hsa-mir-24	dbDEMC
(11)	hsa-mir-193b	dbDEMC
(12)	hsa-mir-106b	dbDEMC
(13)	hsa-mir-15b	dbDEMC
(14)	hsa-mir-100	dbDEMC
(15)	hsa-mir-101	dbDEMC
(16)	hsa-mir-516a	dbDEMC
(17)	hsa-let-7d	dbDEMC
(18)	hsa-mir-125a	dbDEMC
(19)	hsa-let-7f	dbDEMC
(20)	hsa-let-7i	dbDEMC
(21)	hsa-mir-30c	dbDEMC
(22)	hsa-mir-214	dbDEMC
(23)	hsa-mir-513a	dbDEMC
(24)	hsa-mir-484	dbDEMC
(25)	hsa-mir-98	dbDEMC
(26)	hsa-mir-208b	dbDEMC
(27)	hsa-mir-205	dbDEMC
(28)	hsa-let-7g	dbDEMC
(29)	hsa-mir-615	dbDEMC
(30)	hsa-mir-663a	**Unconfirmed**
(31)	hsa-mir-10a	dbDEMC
(32)	hsa-mir-30b	dbDEMC
(33)	hsa-mir-20b	dbDEMC
(34)	hsa-mir-23b	dbDEMC
(35)	hsa-mir-204	dbDEMC
(36)	hsa-mir-519e	dbDEMC
(37)	hsa-mir-515	dbDEMC
(38)	hsa-mir-130b	dbDEMC
(39)	hsa-mir-296	dbDEMC
(40)	hsa-mir-134	dbDEMC
(41)	hsa-mir-132	dbDEMC
(42)	hsa-mir-520h	dbDEMC
(43)	hsa-mir-128	dbDEMC
(44)	hsa-mir-572	dbDEMC
(45)	hsa-mir-30d	dbDEMC
(46)	hsa-mir-197	dbDEMC
(47)	hsa-mir-151a	dbDEMC
(48)	hsa-mir-654	dbDEMC
(49)	hsa-mir-138	dbDEMC
(50)	hsa-mir-495	dbDEMC

Table 5: The top 50 potential isolated diseases predicted of colorectal neoplasms. Forty-nine of the top 50 colorectal neoplasms-related miRNAs have been confirmed by dbDEMC. miRNA hsa-mir-494, which is ranked 45th, has been confirmed by literature.

Rank	miRNA	Evidence
(1)	hsa-mir-29b	dbDEMC
(2)	hsa-mir-15a	dbDEMC
(3)	hsa-mir-223	dbDEMC
(4)	hsa-mir-29c	dbDEMC
(5)	hsa-mir-106b	dbDEMC
(6)	hsa-let-7d	dbDEMC
(7)	hsa-mir-24	dbDEMC
(8)	hsa-mir-100	dbDEMC
(9)	hsa-mir-214	dbDEMC
(10)	hsa-let-7f	dbDEMC
(11)	hsa-let-7g	dbDEMC
(12)	hsa-let-7i	dbDEMC
(13)	hsa-mir-15b	dbDEMC
(14)	hsa-mir-125a	dbDEMC
(15)	hsa-mir-205	dbDEMC
(16)	hsa-mir-101	dbDEMC
(17)	hsa-mir-30b	dbDEMC
(18)	hsa-mir-30c	dbDEMC
(19)	hsa-mir-192	dbDEMC
(20)	hsa-mir-23b	dbDEMC
(21)	hsa-mir-20b	dbDEMC
(22)	hsa-mir-132	dbDEMC
(23)	hsa-mir-138	dbDEMC
(24)	hsa-mir-193b	dbDEMC
(25)	hsa-mir-302b	dbDEMC
(26)	hsa-mir-296	dbDEMC
(27)	hsa-mir-151a	dbDEMC
(28)	hsa-mir-204	dbDEMC
(29)	hsa-mir-196b	dbDEMC
(30)	hsa-mir-10a	dbDEMC
(31)	hsa-mir-30d	dbDEMC
(32)	hsa-mir-212	dbDEMC
(33)	hsa-mir-128	dbDEMC
(34)	hsa-mir-302a	dbDEMC
(35)	hsa-mir-191	dbDEMC
(36)	hsa-mir-302c	dbDEMC
(37)	hsa-mir-197	dbDEMC
(38)	hsa-mir-629	dbDEMC
(39)	hsa-mir-99b	dbDEMC
(40)	hsa-mir-181c	dbDEMC
(41)	hsa-mir-130b	dbDEMC
(42)	hsa-mir-30e	dbDEMC
(43)	hsa-mir-181d	dbDEMC
(44)	hsa-mir-98	dbDEMC
(45)	hsa-mir-494	**PMID: 25270723**
(46)	hsa-mir-452	dbDEMC
(47)	hsa-mir-365a	dbDEMC
(48)	hsa-mir-32	dbDEMC
(49)	hsa-mir-184	dbDEMC
(50)	hsa-mir-424	dbDEMC

Table 6: The top 10 potential isolated miRNA predicted of hsa-mir-106b. All of the top 10 hsa-mir-106b related diseases have been confirmed by dbDEMC, miR2Disease, or HMDD databases.

Rank	Disease	Evidence
(1)	Carcinoma, hepatocellular	HMDD
(2)	Breast neoplasms	dbDEMC, miR2Disease, HMDD
(3)	Stomach neoplasms	HMDD
(4)	Colorectal neoplasms	dbDEMC, miR2Disease
(5)	Lung neoplasms	dbDEMC, miR2Disease
(6)	Melanoma	dbDEMC, HMDD
(7)	Ovarian neoplasms	dbDEMC, HMDD
(8)	Prostatic neoplasms	HMDD
(9)	Heart failure	miR2Disease, HMDD
(10)	Pancreatic neoplasms	dbDEMC, miR2Disease

4. Discussion

The success of SRMDAP could largely be attributed to several factors. First, SRMDAP is a novel method to predict human miRNA-disease associations. This similarity measurement method does not depend on experimentally supported miRNA-disease associations to calculate the functional similarity of miRNAs and diseases. Thus, overestimation of the predictive accuracy was avoided. In SRMDAP, we proposed a density-based recommender model to integrate miRNA similarity subnetwork and disease similarity subnetwork using experimentally verified miRNA-disease associations. Second, SRMDAP incorporates miRNA-mRNA information, disease-gene information, and experimentally verified miRNA-disease associations. This characteristic improved prediction accuracy. Third, only one parameter was used to balance the contributions from miRNA similarity subnetwork and disease similarity subnetwork, and this parameter was easy to adjust. Fourth, LOOCV experiment and case studies about kidney and colorectal neoplasms demonstrated that SRMDAP had excellent predictive performance. Finally, the SRMDAP could predict isolated diseases and isolated miRNAs for disease similarity, and miRNA similarity was obtained independently on the known miRNA-disease associations.

Although SRMDAP contains several innovative concepts, this process has several limitations in its current version. First, a similarity measurement is of vital importance. Hence, miRNA similarity measurement should use more interaction information of miRNAs with other biomolecules. Disease similarity measurement should consider not only functional similarities but also semantic similarities. A fusion of more information sources can benefit the similarity measurement. Second, considering that the SRMDAP is constructed on the basis of known miRNA-disease associations, the performance of SRMDAP can be improved by obtaining more available experimentally verified miRNA-disease associations.

5. Conclusions

Identifying most promising miRNA-disease associations facilitates biological experimentation to save time and cost. In this work, we developed SRMDAP to predict miRNA-disease associations using established miRNA similarity subnetwork and disease similarity subnetwork based on the SimRank and density-based clustering recommender model. We integrated these similarity networks with known experimentally verified miRNA-disease associations using the density-based clustering recommender model. SRMDAP obtained average AUC of 0.8838 in LOOCV. Case studies of kidney and colorectal neoplasms were evaluated, and 49 miRNAs in the top 50 miRNAs were confirmed. SRMDAP also performed well in predicting isolated diseases and miRNAs. For colorectal neoplasms and hsa-mir-106b, all top 50 predicted miRNAs and all top 10 predicted diseases have been confirmed by dbDEMC, miRCancr, HMDD, or the literature. These results demonstrated that SRMDAP has superior performance over the other tested processes.

Conflicts of Interest

There are no conflicts of interest to declare.

Acknowledgments

This work was supported by the Natural Science Foundation of China under Grant no. 61672223 and the Natural Science Foundation of Hunan Provincial under Grant no. 2016jj4029.

References

[1] R. C. Lee, R. L. Feinbaum, and V. Ambros, "The C. elegans heterochronic gene *lin*-4 encodes small RNAs with antisense complementarity to *lin*-14," *Cell*, vol. 75, no. 5, pp. 843–854, 1993.

[2] B. J. Reinhart, F. J. Slack, M. Basson et al., "The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans," *Nature*, vol. 403, no. 6772, pp. 901–906, 2000.

[3] D. P. Bartel, "MicroRNAs: genomics, biogenesis, mechanism, and function," *Cell*, vol. 116, no. 2, pp. 281–297, 2004.

[4] A. M. Cheng, M. W. Byrom, J. Shelton, and L. P. Ford, "Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis," *Nucleic Acids Research*, vol. 33, no. 4, pp. 1290–1297, 2005.

[5] E. A. Miska, "How microRNAs control cell division, differentiation and death," *Current Opinion in Genetics & Development*, vol. 15, no. 5, pp. 563–568, 2005.

[6] X. Karp and V. Ambros, "Encountering microRNAs in cell fate signaling," *Science*, vol. 310, no. 5752, pp. 1288-1289, 2005.

[7] G. A. Calin and C. M. Croce, "MicroRNA signatures in human cancers," *Nature Reviews Cancer*, vol. 6, no. 11, pp. 857–866, 2006.

[8] I. Alvarez-Garcia and E. A. Miska, "MicroRNA functions in animal development and human disease," *Development*, vol. 132, no. 21, pp. 4653–4662, 2005.

[9] L. He, J. M. Thomson, M. T. Hemann et al., "A microRNA polycistron as a potential human oncogene," *Nature*, vol. 435, no. 7043, pp. 828–833, 2005.

[10] M. Lu, Q. Zhang, M. Deng et al., "An analysis of human microRNA and disease associations," *PLoS ONE*, vol. 3, no. 10, Article ID e3420, 2008.

[11] Q. Jiang, Y. Wang, Y. Hao et al., "miR2Disease: a manually curated database for microRNA deregulation in human disease," *Nucleic Acids Research*, vol. 37, no. 1, pp. D98–D104, 2009.

[12] Z. Yang, F. Ren, C. Liu et al., "dbDEMC: a database of differentially expressed miRNAs in human cancers," *BMC Genomics*, vol. 11, supplement 4, article S5, 2010.

[13] B. Xie, Q. Ding, H. Han, and D. Wu, "miRCancer: a microRNA-cancer association database constructed by text mining on literature," *Bioinformatics*, vol. 29, no. 5, pp. 638–644, 2013.

[14] A. Ruepp, A. Kowarsch, D. Schmidl et al., "PhenomiR: a knowledgebase for microRNA expression in diseases and biological processes," *Genome Biology*, vol. 11, no. 1, article R6, 2010.

[15] Z. Yang, L. Wu, A. Wang et al., "DbDEMC 2.0: Updated database of differentially expressed miRNAs in human cancers," *Nucleic Acids Research*, vol. 45, no. 1, pp. D812–D818, 2017.

[16] Q. Zou, J. Li, L. Song, X. Zeng, and G. Wang, "Similarity computation strategies in the microRNA-disease network: a survey," *Briefings in Functional Genomics*, vol. 15, no. 1, pp. 55–64, 2016.

[17] X. Zeng, X. Zhang, and Q. Zou, "Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks," *Briefings in Bioinformatics*, vol. 17, no. 2, pp. 193–203, 2016.

[18] X. Chen, D. Xie, Q. Zhao, and Z.-H. You, "MicroRNAs and complex diseases: from experimental results to computational models," *Briefings in Bioinformatics*, bbx130, 2017.

[19] Q. Jiang, Y. Hao, G. Wang et al., "Prioritization of disease microRNAs through a human phenome-microRNAome network," *BMC Systems Biology*, vol. 4, supplement 1, article S2, 2010.

[20] D. Wang, J. Wang, M. Lu, F. Song, and Q. Cui, "Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases," *Bioinformatics*, vol. 26, no. 13, pp. 1644–1650, 2010.

[21] X. Chen, M. X. Liu, and G. Y. Yan, "RWRMDA: predicting novel human microRNA-disease associations," *Molecular BioSystems*, vol. 8, no. 10, pp. 2792–2798, 2012.

[22] P. Xuan, K. Han, and M. Guo, "Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors," *PLoS ONE*, vol. 8, no. 8, Article ID e70204, 2013.

[23] X. Chen, C. C. Yan, X. Zhang et al., "WBSMDA: within and between score for MiRNA-disease association prediction," *Scientific Reports*, vol. 6, article 21106, 2016.

[24] Q. Zou, J. Li, Q. Hong et al., "Prediction of microRNA-disease associations based on social network analysis methods," *BioMed Research International*, vol. 2015, Article ID 810514, 9 pages, 2015.

[25] C. Gu, B. Liao, X. Li, and K. Li, "Network consistency projection for human miRNA-disease associations inference," *Scientific Reports*, vol. 6, Article ID 36054, 2016.

[26] X. Li, Y. Lin, and C. Gu, "A network similarity integration method for predicting microRNA-disease associations," *RSC Advances*, vol. 7, no. 51, pp. 32216–32224, 2017.

[27] Z.-H. You, Z.-A. Huang, Z. Zhu et al., "PBMDA: a novel and effective path-based computational model for miRNA-disease association prediction," *PLoS Computational Biology*, vol. 13, no. 3, Article ID e1005455, 2017.

[28] C. Gu, B. Liao, X. Li et al., "Network-based collaborative filtering recommendation model for inferring novel disease-related miRNAs," *RSC Advances*, vol. 7, no. 71, pp. 44961–44971, 2017.

[29] Q. Jiang, G. Wang, S. Jin, Y. Li, and Y. Wang, "Predicting human microRNA-disease associations based on support vector machine," *International Journal of Data Mining and Bioinformatics*, vol. 8, no. 3, pp. 282–293, 2013.

[30] X. Chen and G.-Y. Yan, "Semi-supervised learning for potential human microRNA-disease associations inference," *Scientific Reports*, vol. 4, article 5501, 2014.

[31] J. Xu, C.-X. Li, J.-Y. Lv et al., "Prioritizing candidate disease miRNAs by topological features in the miRNA target-dysregulated network: case study of prostate cancer," *Molecular Cancer Therapeutics*, vol. 10, no. 10, pp. 1857–1866, 2011.

[32] J.-Q. Li, Z.-H. Rong, X. Chen, G.-Y. Yan, and Z.-H. You, "MCMDA: matrix completion for MiRNA-disease association prediction," *Oncotarget* , vol. 8, pp. 21187–21199, 2017.

[33] X. Chen, C. Clarence Yan, X. Zhang et al., "RBMMMDA: predicting multiple types of disease-microRNA associations," *Scientific Reports*, vol. 5, article 13877, 2015.

[34] X. Chen, Y. Gong, D. H. Zhang, Z. H. You, and Z. W. Li, "DRMDA: deep representations-based miRNAdisease association prediction," *Journal of Cellular and Molecular Medicine*, vol. 22, no. 1, pp. 472–485, 2018.

[35] Q. Xiao, J. Luo, C. Liang, J. Cai, and P. Ding, "A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations," *Bioinformatics*, vol. 34, no. 2, pp. 239–248, 2018.

[36] Y. Liu, X. Zeng, Z. He, and Q. Zou, "Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources," *IEEE Transactions on Computational Biology and Bioinformatics*, vol. 14, no. 4, pp. 905–915, 2017.

[37] Q. Jiang, G. Wang, and Y. Wang, "An approach for prioritizing disease-related microRNAs based on genomic data integration," in *Proceedings of the 3rd International Conference on BioMedical Engineering and Informatics (BMEI '10)*, pp. 2270–2274, IEEE, October 2010.

[38] H. Shi, G. Zhang, M. Zhou et al., "Integration of multiple genomic and phenotype data to infer novel miRNA-disease associations," *PLoS ONE*, vol. 11, no. 2, Article ID e0148521, 2016.

[39] G. Jeh and J. Widom, "SimRank: a measure of structural-context similarity," in *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 538–543, July 2002.

[40] M. Ester, H.-P. Kriegel, and X. Xu, "Density-based clustering in spatial databases: the algorithm GDBSCAN and its applications," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 169–194, 1998.

[41] S. D. Hsu, Y. T. Tseng, S. Shrestha et al., "miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions," *Nucleic Acids Research*, vol. 42, no. D1, pp. D78–D85, 2014.

[42] J. Piñero, N. Queralt-Rosinach, À. Bravo et al., "DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes," *Database*, vol. 2015, Article ID bav028, 2015.

[43] Y. Li, C. Qiu, J. Tu et al., "HMDD v2.0: a database for experimentally supported human microRNA and disease associations," *Nucleic Acids Research*, vol. 42, no. D1, pp. D1070–D1074, 2014.

[44] Y. Pi, H. Peng, S. Zhou, and Z. Zhang, "A scalable approach to column-based low-rank matrix approximation," in *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pp. 1600–1606, August 2013.

[45] R. Meymandpour and J. G. Davis, "A semantic similarity measure for linked data: An information content-based approach," *Knowledge-Based Systems*, vol. 109, pp. 276–293, 2016.

[46] J. Gao, F. Liang, W. Fan, C. Wang, Y. Sun, and J. Han, "On community outliers and their efficient detection in information networks," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 813–822, ACM, Washington, DC, USA, July 2010.

[47] H. Cheng, Y. Zhou, and J. X. Yu, "Clustering large attributed graphs: a balance between structural and attribute similarities," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 5, no. 2, article 12, 33 pages, 2011.

[48] P. Lee, L. V. S. Lakshmanan, and J. X. Yu, "On top-k structural similarity search," in *Proceedings of the IEEE 28th International Conference on Data Engineering (ICDE '12)*, pp. 774–785, April 2012.

[49] S. Bandyopadhyay, R. Mitra, U. Maulik, and M. Q. Zhang, "Development of the human cancer microRNA network," *Silence*, vol. 1, no. 1, article 6, 2010.

[50] T. Nepusz, H. Yu, and A. Paccanaro, "Detecting overlapping protein complexes in protein-protein interaction networks," *Nature Methods*, vol. 9, no. 5, pp. 471-472, 2012.

[51] SEER Stat Fact Sheets: Kidney and Renal Pelvis Cancer, National Cancer Institue.

[52] R. M. Silva-Santos, P. Costa-Pinheiro, A. Luis et al., "microRNA profile: a promising ancillary tool for accurate renal cell tumour diagnosis," *British Journal of Cancer*, vol. 109, no. 10, pp. 2646–2653, 2013.

[53] Z. Yu, L. Ni, D. Chen et al., "Identification of miR-7 as an oncogene in renal cell carcinoma," *Journal of Molecular Histology*, vol. 44, no. 6, pp. 669–677, 2013.

[54] H. Brenner, M. Kloor, and C. P. Pox, "Colorectal cancer," *The Lancet*, vol. 383, no. 9927, pp. 1490–1502, 2014.

[55] Y. Okugawa, W. M. Grady, and A. Goel, "Epigenetic alterations in colorectal cancer: emerging biomarkers," *Gastroenterology*, vol. 149, no. 5, pp. 1204–1225.e12, 2015.

[56] A. Yamada, T. Horimatsu, Y. Okugawa et al., "Serum MIR-21, MIR-29a, and MIR-125b are promising biomarkers for the early detection of colorectal neoplasia," *Clinical Cancer Research*, vol. 21, no. 18, pp. 4234–4242, 2015.

[57] H.-B. Sun, X. Chen, H. Ji et al., "MiR-494 is an independent prognostic factor and promotes cell migration and invasion in colorectal cancer by directly targeting PTEN," *International Journal of Oncology*, vol. 45, no. 6, pp. 2486–2494, 2014.

Hindawi
BioMed Research International
Volume 2018, Article ID 3195025, 10 pages
https://doi.org/10.1155/2018/3195025

Research Article

Suppression of *IL-6* Gene by shRNA Augments Gemcitabine Chemosensitization in Pancreatic Adenocarcinoma Cells

Hai-Bo Xing,[1] Meng-Ting Tong,[2] Jing Wang,[2] Hong Hu,[2] Chong-Ya Zhai,[2] Chang-Xin Huang,[3] and Da Li [2]

[1]*Department of ICU, Sir Run Run Shaw Hospital, School of Medicine, Zhejiang University, Hangzhou, China*
[2]*Department of Medical Oncology, Sir Run Run Shaw Hospital, School of Medicine, Zhejiang University, Hangzhou, China*
[3]*Department of Medical Oncology, The Affiliated Hospital of Hangzhou Normal University, Zhejiang University, Hangzhou, China*

Correspondence should be addressed to Da Li; lidaonconew@zju.edu.cn

Received 16 October 2017; Revised 19 January 2018; Accepted 29 January 2018; Published 6 March 2018

Academic Editor: Tao Huang

Pancreatic adenocarcinoma has an exceedingly poor prognosis, accounting for five-year survival of less than 5%. Presently, improving the efficacy of pancreatic adenocarcinoma treatment has been the focus of medical researchers worldwide. Recently, it has been suggested that deregulation of interleukin- (IL-) 6 is caused by a key gene involved in the beginning and development of pancreatic adenocarcinoma. Herein, we investigated whether suppression of IL-6 could augment gemcitabine sensitivity in the PANC-1 cells. We found considerably higher expression of IL-6 in pancreatic adenocarcinoma tissues than that in the adjacent nontumorous tissues. Suppression of IL-6 by shRNA resulted in apoptosis as well as inhibition of cell proliferation and tumorigenicity. In addition, suppression of IL-6 remarkably promoted antitumor effect of gemcitabine, indicating that the combination of shRNA targeting IL-6 with gemcitabine may provide a potential clinical approach for pancreatic cancer therapy.

1. Introduction

Pancreatic adenocarcinoma is a type of malignant tumor characterized by an extremely low 5-year survival of 5% or less [1, 2]. Local invasion and lymphatic metastasis occur in 80% of the clinical cases, and their prognosis of disease progression is poor [3, 4]. While radiotherapy is one of the most important therapeutic methods for pancreatic adenocarcinoma, chemotherapy with gemcitabine following tumor surgery can contribute remarkably to a substantial delay of relapse. After absorption by cells, gemcitabine is phosphorated by deoxycytidine kinase to form dFdCTP, which competes with dCTP for insertion into the deoxycytidine sites of DNA strands, by which mechanism it destructs DNA replication and consequently results in cell death. However, the functional mechanism of gemcitabine is still obscure, and further studies are necessary to elucidate the mechanism of gemcitabine resistance.

Interleukin-6 (IL-6), a cytokine with pleiotropic function, is synthesized by various types of cells, such as endothelial cells, macrophages, myeloid cells, and fibroblasts. It promotes cell proliferation and differentiation, participates in the immune defense system, and is involved in almost all physiological processes. Moreover, the involvement of IL-6 is critically important in the development of various conditions, such as infections, traumas, and hematopoiesis [5–8]. The expression of IL-6 has been found to increase dramatically in cases of a number of different cancer types, including liver, colon, and ovarian cancer [9–11]. In pancreatic adenocarcinoma patients, IL-6 upregulates other cytokine expression and is involved in proliferation, resistance to apoptosis, and immune evasion [12]. IL-6 becomes bound to a composite receptor consisting of a distinct 80 kDa alpha subunit (IL-6Rα, gp80) and two gp130 receptor subunits transducing a signal after the ligand binding. Then, gp130 triggers the activation of Janus kinase (JAK2)/signal transducer and activator of transcription 3 (STAT3) pathway through the induction of cross-phosphorylation of the JAK2 peptides linked to the cytoplasmic side of the receptor [13, 14]. STAT3 is an important member of the STAT family that is constitutively

induced in a broad spectrum of human malignant cancers [15] and is recognized as a major factor for oncogenesis in many epithelial malignant tumors. STAT3 is also tightly involved in the development of skin and gastric cancers in mouse models [16, 17]. Considering the substantially important roles of STAT3 in the development and progression of a large number of cancer types, we speculated that blocking IL-6 expression in STAT3 signaling may be valuable in the treatment of pancreatic adenocarcinoma.

In this study, we demonstrated that suppression of IL-6 expression resulted in inhibition of *in vitro* and *in vivo* tumorigenicity of the pancreatic adenocarcinoma cells. In addition, knockdown of IL-6 expression sensitized pancreatic adenocarcinoma cells to gemcitabine. These findings not only revealed the vital role IL-6 exerts in pancreatic adenocarcinoma development, but also highlighted that the combination of gemcitabine and shRNA targeting IL-6 can potentially be applied in clinical practice for the treatment of pancreatic adenocarcinoma.

2. Materials and Methods

2.1. Plasmids and Cell Culture. IL-6 shRNA expressing plasmids were designed and constructed by Genechem Corporation (Shanghai, China). The sequence of shRNA-IL-6 was "gacactattttaattatttttaa."

Dulbecco's modified Eagle's medium (DMEM)/F12 supplemented with 10% fetal bovine serum (FBS) was used for culturing all cell lines. To suppress IL-6 expression, a plasmid containing IL-6-shRNA was transfected into the PANC-1 cells, and the positive clones were identified using puromycin (300 ng/mL) for a period of 24 h.

2.2. MTT Assay. Effect of shIL-6, gemcitabine, or shIL-6 transfection plus gemcitabine treatment on the proliferation of PANC-1 cells was determined by MTT assay. Briefly, a volume of 20 μL MTT was applied to each well of a 96-well plate and incubated with the cells at 37°C for 4 h. Next, 150 μL/well DMSO was admixed to obtain coloration, and samples were shaken gently for 2 h at room temperature in the dark. To measure cell viability, the values of absorbance at 490 nm were determined. Data from triplicate wells per treatment were collected and the experiments were repeated three times at separate occasion.

2.3. Assessment of Colony Formation. The transfected PANC-1 cells were transferred into a six-well plate and were cultured in DMEM with fetal bovine serum (10%). After 24 h, the medium was replaced with fresh medium containing G418 (400 mg/mL) and changed every other day. After 14-day incubation, cells were fixed with methanol and stained with 0.1% crystal violet. The number of visible colonies was counted.

2.4. Evaluation of Cell Invasion. Assessment of cell invasion was conducted with the Boyden chambers using coated Matrigel following the instruction provided by the manufacturer (BD Biosciences, San Jose, CA). Staining of the invasive cancer cells was performed with crystal violet, and they were observed and counted under a microscope. The experiments were performed twice in triplicate at minimum.

2.5. Flow Cytometry Analysis. Detection of cellular apoptosis was done by double staining of Annexin V and propidium iodide (PI) following the previously published method [18]. Trypsin (0.25%) was utilized to harvest the cells, followed by double washing with PBS. Further, the cells were resuspended in binding buffer (250 μL) that was calibrated to 1×10^6/mL. Then, to the cell suspension, the staining solution that contained Annexin V-FITC and PI was added followed by incubation for 30 min in the dark. Apoptosis was then analyzed by flow cytometry using FACSAria System (Becton Dickinson, Franklin Lakes, NJ, USA).

2.6. Quantitative Real Time RT-PCR. Total RNA was extracted with the RNeasy Mini kit (Qiagen) following the manufacturer's guidelines. The specific experiments were conducted with SYBR Green Power Master Mix complying with the instructions provided by the manufacturer (Applied Biosystems). Reverse transcription of the reaction mixture (20 μL) containing total RNA (1 μg) to cDNA was done with PrimeScript RT-polymerase (Life, Shanghai, China). Then, the cDNA was used for real time quantitative PCR utilizing primers (Generay, Shanghai, China) distinct for STAT3, JAK2, Bcl-2, Bax, caspase-3, and caspase-9. GAPDH was used as an internal control. All reactions were carried out on the Applied Biosystems 7500 Sequence Detection System (Applied Biosystems, Foster City, CA, USA). The levels of relative expression were calculated as ratios normalized against those of GAPDH. Comparative quantification was performed using the $2^{-\Delta\Delta Ct}$ method. Sequences of the primers were listed in Table 1.

2.7. Immunoblotting. Denaturing SDS-PAGE sample buffer was utilized for PANC-1 cell lysis through standard methods. Then, separation of protein lysates was done with 10% SDS-PAGE, and they were transferred onto nitrocellulose membranes. TBS containing 0.1% Triton X-100 and 5% nonfat milk was used overnight at 4°C to block the membranes, followed by incubation at 4°C overnight with primary antibodies of STAT3, p-STAT3, JAK, p-JAK, Bcl-2, Bax, caspase-3, caspase-9, and β-actin (Santa Cruz Biotech, Santa Cruz, CA, USA). Upon washing, the membranes were subjected to incubation with HRP-conjugated secondary antibodies for 2 h at room temperature. Detection of the signal was accomplished with the ECL reagents.

2.8. Animal Experiments. The animal experiments were conducted in compliance with the Guide for the Care and Use of Laboratory Animals, and the study protocol was approved by the Animal Ethics Committee of Zhejiang University. Injection of 1×10^7 cells in 100 μL PBS was subcutaneously administered in the right back area of nude mice. Then, upon reaching tumor volume of approximately 150 mm^3, the animals were randomly divided into groups (20 animals each group) of nontreated control (NC), intraperitoneally injected

Table 1: Sequences of the primers for PCR.

Targets	Direction	Sequence
Bcl-2	Forward	GCCTTCTTTGAGTTCGGTG
	Reverse	AGTCATCCACAGGGCGAT
Bax	Forward	ATGGGCTGGACATTGGACTT
	Reverse	GCCACAAAGATGGTCACGGT
Caspase-3	Forward	ATCCAGTCGCTTTGTGCCAT
	Reverse	TTCTGTTGCCACCTTTCGGT
Caspase-9	Forward	TGGGCTCACTCTGAAGACCT
	Reverse	AGCAACCAGGCATCTGTTTA
JAK2	Forward	GCCTTCTTTCAGAGCCATCA
	Reverse	CCAGGGCACCTATCCTCATA
STAT3	Forward	AATACCATTGACCTGCCGATGT
	Reverse	GGTGGTCTCCTCTGACTTCAACA
GAPDH	Forward	GGTGGTCTCCTCTGACTTCAACA
	Reverse	GTTGCTGTAGCCAAATTCGTTGT

gemcitabine alone at a dose of 100 mg/kg daily, injection of shIL-6-plasmid, or injection of shIL-6 plus gemcitabine. Indicators of drug toxicity, including behavior shifts, loss of weight, and changed feeding patterns were incessantly monitored during the entire trial. The tumors were collected from 4 animals of each treatment group and weighed at weeks 1, 2, 3, 4, and 5, and the tumor masses were established.

2.9. Immunohistochemical Staining. Streptavidin-peroxidase (SP) staining was performed for protein detection after retrieval of the antigens through microwave treatment. After inhibition of the activity of endogenous peroxidase by incubation in 3% H_2O_2 for 10 min, the samples were washed with PBS and incubated with anti-p-STAT3 or p-JAK2 antibodies at 4°C overnight (PBS served as a negative control). After coculture of the specimens with the respective secondary antibodies for 30 min, the specimens were subjected to streptavidin-peroxidase treatment for additional 30 min. After rinsing with PBS, diaminobenzidine (DAB) solution was applied. Next, counterstaining with hematoxylin was performed.

2.10. Statistical Analysis. SPSS15.0 software was used for the general statistical analysis. Student's t-test, one-way analysis of variance (ANOVA), χ^2 test, or Wilcoxon test was employed to appropriately assess the significance between groups. All tests performed were two-sided, and $P < 0.05$ was considered statistically significant.

3. Results

3.1. Elevated IL-6 Expression Was Associated with Pancreatic Adenocarcinoma. Recent studies indicated the existence of a correlation between IL-6 expression and cancer cell. To identify the association of IL-6 expression and the development of pancreatic adenocarcinoma, by using real time RT-PCR and immunoblot, we compared the IL-6 mRNA and protein levels in pancreatic adenocarcinoma with that in their adjacent nontumor tissues. We found that the IL-6 expression in the pancreatic adenocarcinoma was significantly higher than that in the adjacent nontumor tissues (Figures 1(a), 1(b), and 1(c)).

3.2. Suppression of IL-6 and Its Effect on Gemcitabine-Mediated Antiproliferation and Apoptosis Induction. To examine whether IL-6 affects tumor cell proliferation and apoptosis, we established stable transfectants with shIL-6 expression in PANC-1 cells. We found that IL-6 expression was significantly downregulated after 24 hours of transfection of shIL-6 into the cells. To determine the effect of shIL-6 on cell proliferation, the effects of the of shIL-6 alone or shIL-6 plus 100 μM gemcitabine on the PANC-1 cells were examined by MTT assays. We observed that shIL-6 significantly inhibited PANC-1 cell proliferation in a time-dependent manner. Gemcitabine dramatically enhanced the inhibitory effect of shIL-6 (Figure 2). Furthermore, to evaluate the shIL-6-mediated promotion of apoptosis, PANC-1 cells with stably expressing control vector or shIL-6 and shIL-6 combined with gemcitabine were subjected to Annexin V analysis by flow cytometry. Suppression of IL-6 by shIL-6 alone resulted in significant apoptosis of the cells (20.2% versus 5,2% of control, Figure 3(b)), which was further augmented by the addition of gemcitabine (41.2%, Figure 3(d)).

3.3. Effect of IL-6 Suppression on Gemcitabine-Mediated Inhibition of Colony Formation and Cell Invasion. We measured the *in vitro* tumorigenicity of the PANC-1 cells using a soft-agar assay and found that PANC-1 cells treated with gemcitabine alone or transfected with shIL-6 formed fewer colonies than the cells transfected with control vector. As expected, combination of shIL-6 transfection and gemcitabine treatment led to a more pronounced decrease in colony formation capacity compared to other groups (Figures

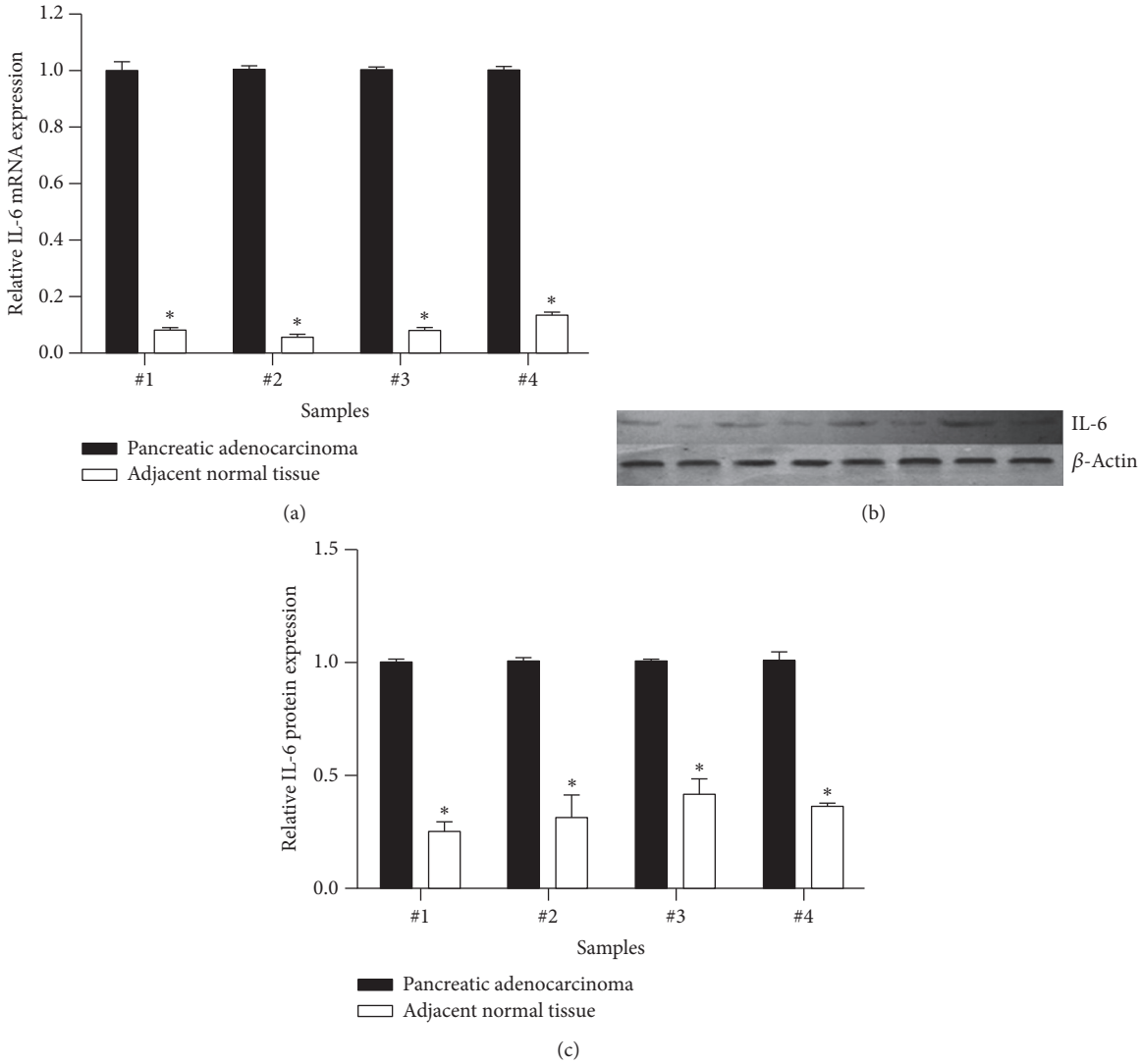


Figure 1: Expression of IL-6 mRNA and protein in four pancreatic adenocarcinoma tissues and their adjacent nontumorous tissues. Expression of IL-6 mRNA was quantified by real time RT-PCR and protein was semiquantitatively assessed by immunoblotting as described in the methods. $^{*}P < 0.05$ compared with pancreatic adenocarcinoma.

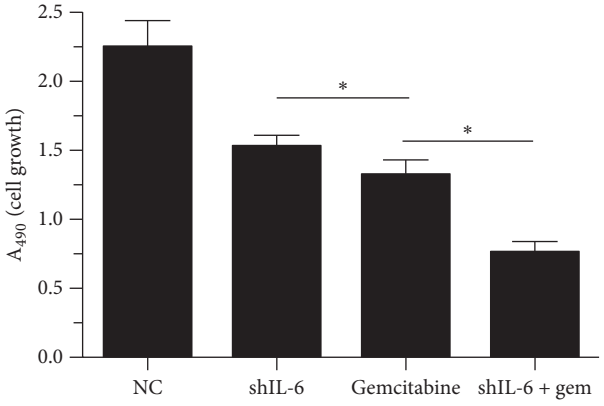


Figure 2: Effect of shIL-6 transfection and gemcitabine treatment on proliferation of PANC-1 cells. MTT assay was performed in the PANC-1 cells 48 h after shIL-6 transfection followed by treatment with or without gemcitabine. $^{*}P < 0.05$.

4(a) and 4(b)). To elucidate whether the knockdown of IL-6 could enhance gemcitabine-mediated tumor cell invasion, an invasion assay was performed. The results revealed that gemcitabine induced an approximately 3-fold decrease, that shIL-6 induced nearly a 2-fold reduction in cell invasion, and that the treatment with gemcitabine in the cells transfected with shIL-6 caused an even more remarkable decrease of cell invasion (7-fold) (Figures 4(c) and 4(d)).

3.4. Combined Effect of shIL-6 Transfection and Gemcitabine Treatment on Signal Transduction and Apoptosis-Associated Proteins. To elucidate the mechanism of the combination of shIL-6 and gemcitabine on apoptosis, expression of STAT3, p-STAT3, JAK2, p-JAK2 Bcl-2, Bax, caspase-3, and caspase-9 was investigated. After 24 h transfection of shIL-6, the PANC-1 cells were treated with or without 100 μM gemcitabine. GAPDH or β-actin was utilized as an internal control. Real

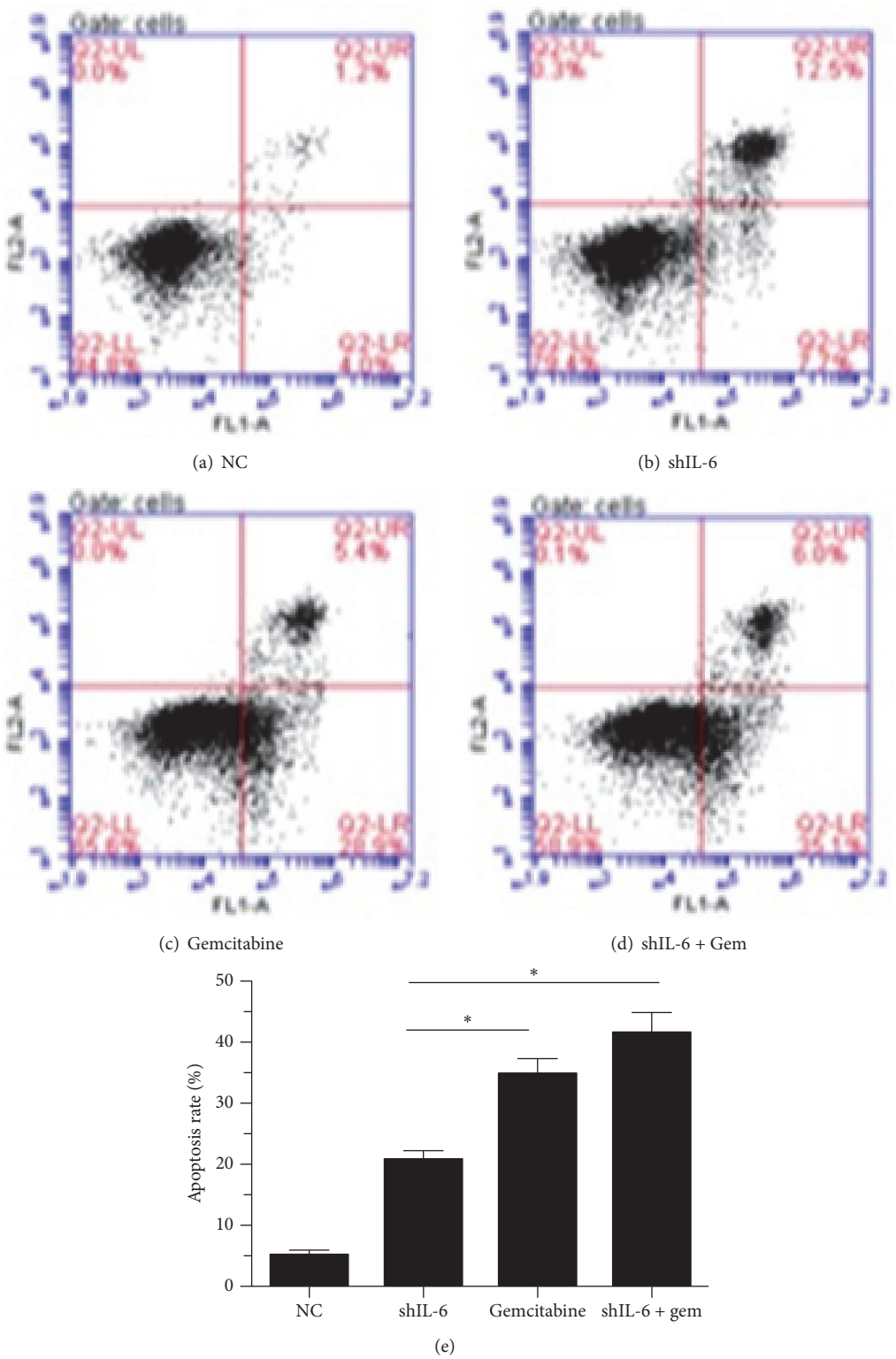


Figure 3: Effect of shIL-6 transfection and gemcitabine treatment on apoptosis of PANC-1 cells. Apoptosis assay was performed in the PANC-1 cells by flow cytometry analysis of Annexin V/PI staining 48 h after shIL-6 transfection followed by treatment with or without gemcitabine. ((a)–(d)) Representative results of flow cytometry analysis. (e) An average of 3 separate experiments, $^{*}P < 0.05$.

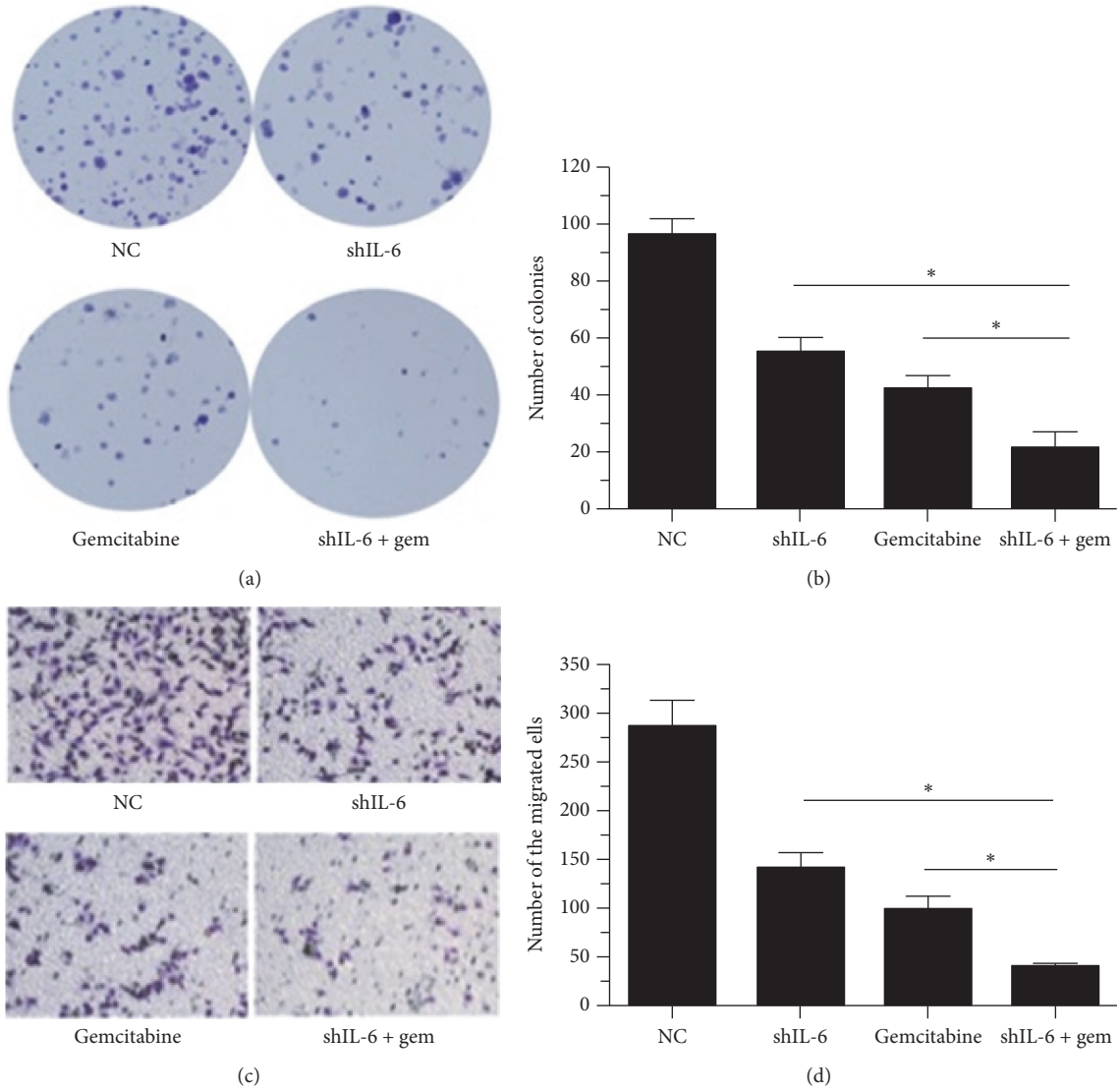


Figure 4: Effect of IL-6 suppression on colony formation and cell invasion. ((a) and (b)) Soft-agar assay was performed in the cells with stably expressing control vector or the shIL-6 and treated with or without 100 mg/kg gemcitabine. Data presented were from three independent experiments with duplicates in each experiment. ((c) and (d)) Invasiveness of the cells with stably expressing control vector or the shIL-6, treated with or without gemcitabine. Cellular invasive capability was determined with a modification of the Boyden chamber invasion assay as described in the Materials and Methods and representative images were presented in (c). The percentage of invasive cells from three separate experiments (mean ± SD) was presented as bar graphs (d). $^{*}P < 0.05$.

time RT-qPCR and Western blots results were performed in three separate experiments. Expression of the genes was normalized to the control in the respective experiment, and a value of 1.0 was assigned. In the group treated with the combination of shRNA-IL-6/gemcitabine (100 μM), the expression levels of STAT3, p-STAT3, JAK, p-JAK, and Bcl-2 were substantially lower than those of the group with the single treatment. On the other hand, the expression levels of Bax, caspase-3, and caspase-9 in both the single treatment group and the group of the treatment with the shRNA-IL-6/gemcitabine (100 μM) combination were higher than the ones in the control treatment (Figures 5(a), 5(b), and 5(c)).

3.5. The Combination of Gemcitabine and shIL-6 Enhanced an In Vivo Antitumor Effect. To examine the *in vivo* tumorigenicity, we performed xenograft experiments, in which we administered injections in the right back area of nude mice with the following content: (1) PANC-1 cells with the control vector, (2) control vector with 10 mg/kg gemcitabine daily for 10 d via *i.p.* injection, (3) shIL-6, and (4) shIL-6 combined with 10 mg/kg gemcitabine daily for 10 d via *i.p.* injection. Mice were injected with 1×10^7 of cells. As shown in Figure 6, gemcitabine or suppression of IL-6 caused a significant reduction in the size compared to the vector control group. Interestingly, the application of gemcitabine combined with

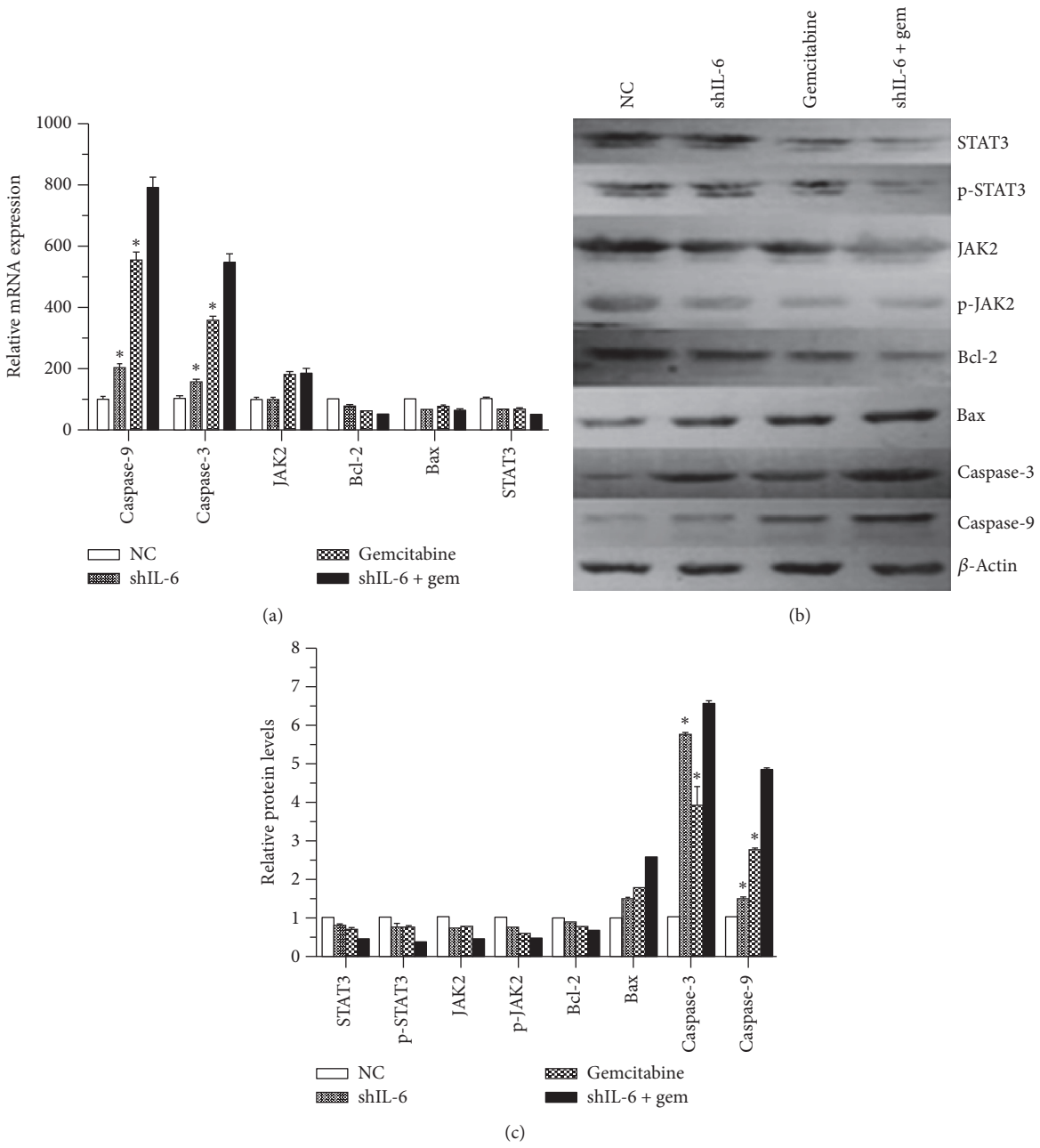


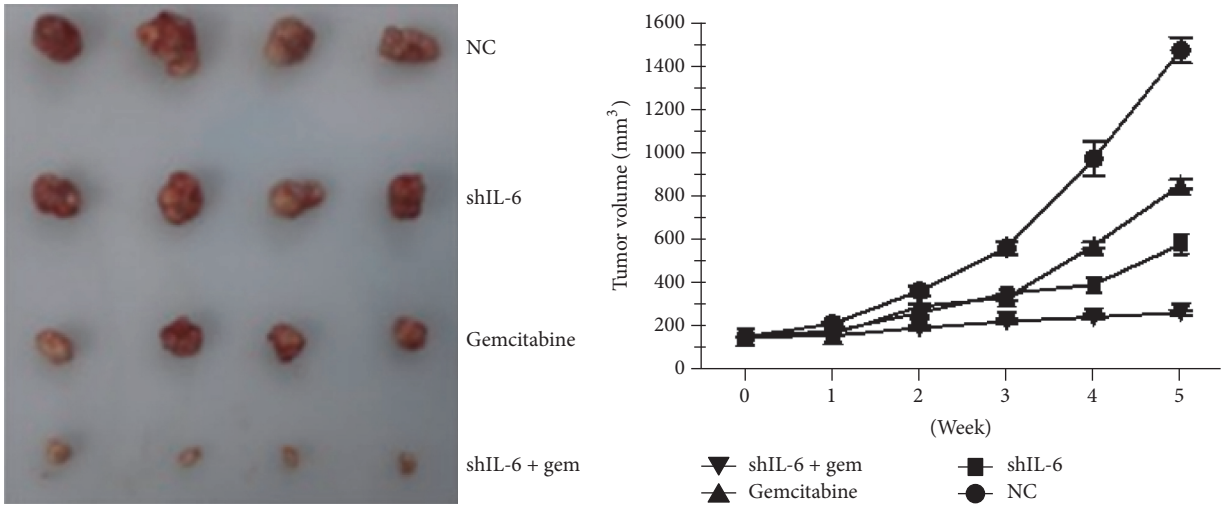
Figure 5: Effect of IL-6 suppression on STAT3 signaling and apoptosis-associated proteins. (a) Effect on the expression of caspase-3, caspase-9, IL-6, STAT3, JAK2, and Bcl2 mRNA. Expression of the indicated mRNA was quantified by real time RT-PCR in the cells after 48 hours of shIL-6 transfection followed by treatment with or without gemcitabine. $^{*}P < 0.05$ compared with the nontreated control (NC). ((b) and (c)) Protein levels of STAT3, p-STAT3, JNK, p-JNK, Bcl-2, Bax, caspase-3, and caspase-9. Levels of the indicated proteins were semiquantitatively determined by immunoblot in the cells 48 hours after shIL-6 transfection followed by treatment with or without gemcitabine. Representative image data (b) as well as an average of three separate experiments (c) were presented. $^{*}P < 0.05$ compared with shIL-6 + gemcitabine.

shIL-6 resulted in an even more dramatic decrease of tumor size and weight (Figures 6(a) and 6(b)).

We then detected the protein expression of p-STAT3 and p-JAK2 in the different tumor groups by immunohistochemical staining. As illustrated in Figure 7, the immunostaining intensity of p-STAT3 and p-JAK2 in shRNA-IL-6 alone or gemcitabine (100 μM) alone groups was lower than that in the control group, and the intensity was even lower in the treatment with the combination of shRNA-IL-6/gemcitabine (100 μM).

4. Discussion

Many research results have evidenced that the one-year survival rate of pancreatic adenocarcinoma patients is less than 20%, and the five-year survival rate is under 5%.



(a) (b)

Figure 6: Antitumor activity of shIL-6 and gemcitabine in the xenograft model. Nude mice were randomly assigned to the following four groups: PANC-1 cells with stably expressing control vector, control vector with 100 mg/kg gemcitabine daily for 10 d via *i.p.* injection, shIL-6 alone, and shIL-6 plus 10 mg/kg gemcitabine daily for 10 d via *i.p.* injection. (a) Macrograph of the tumor tissues obtained on week 5. (b) Comparison of the tumor volumes as function of time.

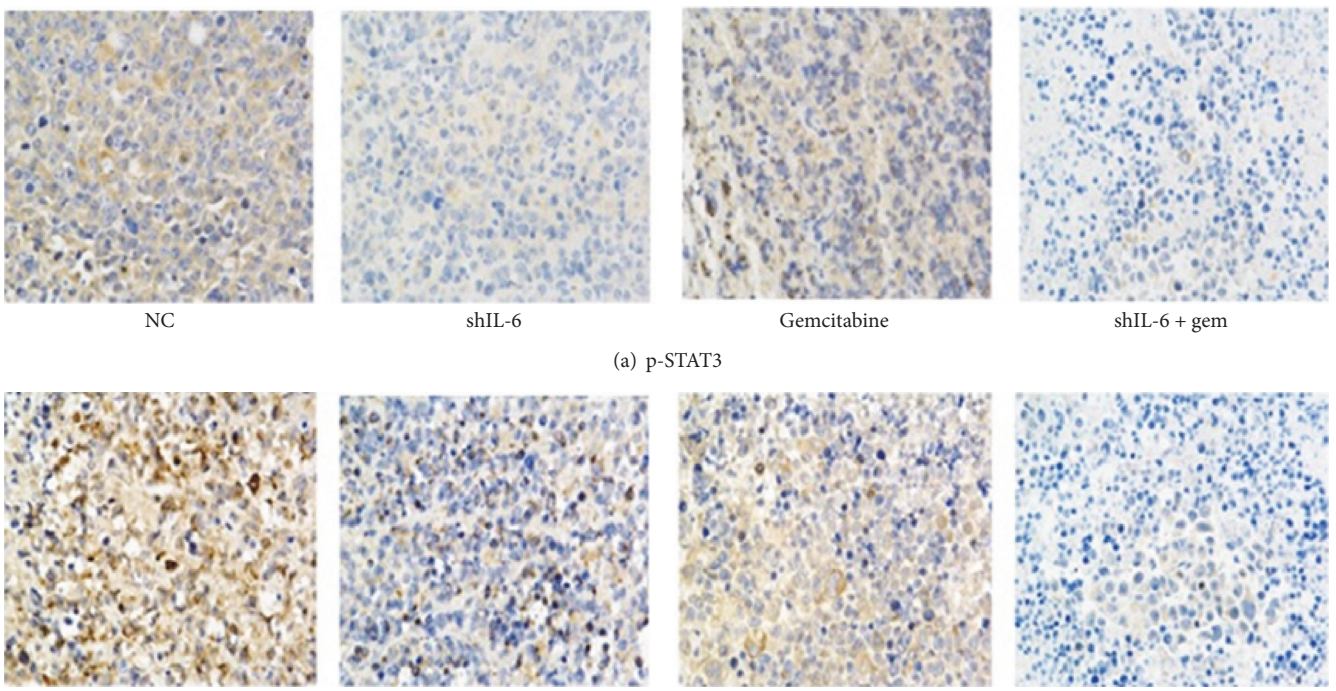


Figure 7: Representative images of p-STAT3 and p-JAK2 expression in the nude mice tumors assessed by immunohistochemistry. Magnification: ×200.

Moreover, wide cancer metastasis has happened in most diagnosed pancreatic adenocarcinoma patients. Eventually, the growth, metastasis, and invasion of cancer cells cause the death of the patients [19]. At present, in spite of the advances in surgical resection, radiotherapy, and chemotherapy, the contemporary therapeutic treatments have not succeeded in significantly improving the survival rate of patients with pancreatic adenocarcinoma [20]. Gemcitabine has been widely used in the treatment of pancreatic adenocarcinoma, but the acquired and intrinsic resistance of pancreatic adenocarcinoma cells can often lead to the need for repeated treatments [21]. The therapeutic targeting of pancreatic adenocarcinoma genes has attracted substantial scientific attention over the recent years [22]. The results of an experiment with mice revealed that the oncogenesis of melanoma and prostatic or gastric cancers could be inhibited by blocking of the IL-6/STAT3 signal pathway [23]. These findings indicated the potential application of interventions in the IL-6/STAT3 pathway for pancreatic adenocarcinoma therapy.

In our study, a shIL-6-plasmid was successfully constructed and used for transfection into the pancreatic adenocarcinoma cell strain PANC-1, achieving a silencing rate of more than 70%. The levels of protein and mRNA expression of STAT3 and p-STAT3 were considerably downregulated by interference with IL-6 expression. Meanwhile, we found that suppression of IL-6 induced apoptosis and reduced the cell survival or tumorigenicity of the cancer cells. These findings not only confirmed the vital role of IL-6 in development of pancreatic adenocarcinoma, but also highlighted that shRNA targeting IL-6 has the potential to become a powerful tool with its low toxicity, remarkable efficiency, and pronounced specificity.

Pancreatic adenocarcinoma cells have an acquired or intrinsic resistance to gemcitabine. In a previous examination, human pancreatic cancer cell lines (BxPc-3, PancTu-1, and Capan-1) were treated with sulfasalazine, an inhibitor of NF-Kβ gene, and it was found that sulfasalazine enhanced sensitivity of the cancer cells to chemotherapy [24]. Effect of the combination of gemcitabine and molecularly targeted agents, such as IL-6 inhibitor (shRNA), has been encouraging. In our study, proliferation of the cancer cells was substantially suppressed in a time-dependent pattern when gemcitabine and shRNA targeting IL-6 were used together. Moreover, this combination remarkably induced PANC-1 cell apoptosis. To better characterize the observed synergic effect, a tumor xenograft study was performed, in which shRNA targeting IL-6 and gemcitabine were used either alone or in combination. The combined treatment dramatically inhibited tumor growth compared to either treatment alone, suggesting that shRNA targeting IL-6 sensitized the xenografted pancreatic adenocarcinoma cells to gemcitabine.

In conclusion, our investigation provides evidence on the significance of shRNA targeting IL-6 in determining the tumorigenicity of pancreatic adenocarcinoma cells. We discovered that sensitivity of PANC-1 cells to gemcitabine could be enhanced by suppression of IL-6 with shRNA. Therefore, the combination of gemcitabine and shRNA targeting IL-6 may become a potential clinical application for the treatment of pancreatic adenocarcinoma.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by Medicine and Health Technology Plan of Zhejiang Province (no. 2015KYB217).

References

[1] M. Carpelan-Holmström, S. Nordling, E. Pukkala et al., "Does anyone survive pancreatic ductal adenocarcinoma? A nationwide study re-evaluating the data of the Finnish Cancer Registry," *Gut*, vol. 54, no. 3, pp. 385–387, 2005.

[2] M. Del Chiaro, R. Segersvärd, M. Löhr, and C. Verbeke, "Early detection and prevention of pancreatic cancer: Is it really possible today?" *World Journal of Gastroenterology*, vol. 20, no. 34, pp. 12118–12131, 2014.

[3] D. Li, K. Xie, R. Wolff, and J. L. Abbruzzese, "Pancreatic cancer," *The Lancet*, vol. 363, no. 9414, pp. 1049–1057, 2004.

[4] H. Papatheodorou, A. D. Papanastasiou, C. Sirinian et al., "Expression patterns of SDF1/CXCR4 in human invasive breast carcinoma and adjacent normal stroma: correlation with tumor clinicopathological parameters and patient survival," *Pathology - Research and Practice*, vol. 210, no. 10, pp. 662–667, 2014.

[5] C. Braconi, E. Swenson, T. Kogure, N. Huang, and T. Patel, "Targeting the IL-6 dependent phenotype can identify novel therapies for cholangiocarcinoma," *PLoS ONE*, vol. 5, no. 12, Article ID e15195, 2010.

[6] H. Isomoto, S. Kobayashi, N. W. Werneburg et al., "Interleukin 6 upregulates myeloid cell leukemia-1 expression through a STAT3 pathway in cholangiocarcinoma cells," *Hepatology*, vol. 42, no. 6, pp. 1329–1338, 2005.

[7] G. Budziński, A. Suszka-Switek, P. Roman et al., "Interleukin-6 concentration in the transgenic pig's liver preserved for 24 hours in Biolasol solution," *Transplantation Proceedings*, vol. 46, no. 8, pp. 2552–2554, 2014.

[8] R. Rasool, I. Ashiq, I. A. Shera, Q. Yousuf, and Z. A. Shah, "Study of serum interleukin (IL) 18 and IL-6 levels in relation with the clinical disease severity in chronic idiopathic urticaria patients of Kashmir (North India)," *Asia Pacific Allergy*, vol. 4, no. 4, p. 206, 2014.

[9] J. Wang, A. Sharma, S. A. Ghamande et al., "Serum protein profile at remission can accurately assess therapeutic outcomes and survival for serous ovarian cancer," *PLoS ONE*, vol. 8, no. 11, Article ID e78393, 2013.

[10] A. Juasook, R. Aukkanimart, T. Boonmars et al., "Tumor-related gene changes in immunosuppressive syrian hamster cholangiocarcinoma," *Pathology & Oncology Research*, vol. 19, no. 4, pp. 785–794, 2013.

[11] A. Mantovani, P. Allavena, A. Sica, and F. Balkwill, "Cancer-related inflammation," *Nature*, vol. 454, no. 7203, pp. 436–444, 2008.

[12] T. Masui, R. Hosotani, R. Doi et al., "Expression of IL-6 receptor in pancreatic cancer: Involvement in VEGF induction," *Anticancer Reseach*, vol. 22, no. 6C, pp. 4093–4100, 2002.

[13] S. A. Jones, S. Horiuchi, N. Topley, N. Yamamoto, and G. M. Fuller, "The soluble interleukin 6 receptor: mechanisms of production and implications in disease," *The FASEB Journal*, vol. 15, no. 1, pp. 43–58, 2001.

[14] K. M. Block, N. T. Hanke, E. A. Maine, and A. F. Baker, "IL-6 stimulates STAT3 and Pim-1 kinase in pancreatic cancer cell lines," *Pancreas*, vol. 41, no. 5, pp. 773–781, 2012.

[15] D. A. Frank, "STAT3 as a central mediator of neoplastic cellular transformation," *Cancer Letters*, vol. 251, no. 2, pp. 199–210, 2007.

[16] K. S. Chan, S. Sano, K. Kiguchi et al., "Disruption of Stat3 reveals a critical role in both the initiation and the promotion stages of epithelial carcinogenesis," *The Journal of Clinical Investigation*, vol. 114, no. 5, pp. 720–728, 2004.

[17] B. J. Jenkins, D. Grail, T. Nheu et al., "Hyperactivation of Stat3 in gp130 mutant mice promotes gastric hyperproliferation and desensitizes TGF-β signaling," *Nature Medicine*, vol. 11, no. 8, pp. 845–852, 2005.

[18] I. Vermes, C. Haanen, H. Steffens-Nakken, and C. Reutellingsperger, "A novel assay for apoptosis Flow cytometric detection of phosphatidylserine expression on early apoptotic cells using fluorescein labelled Annexin V," *Journal of Immunological Methods*, vol. 184, no. 1, pp. 39–51, 1995.

[19] S. Munigala, F. Kanwal, H. Xian, and B. Agarwal, "New diagnosis of chronic Pancreatitis: Risk of missing an underlying pancreatic cancer," *American Journal of Gastroenterology*, vol. 109, no. 11, pp. 1824–1830, 2014.

[20] J. Xiao, G. Li, S. Lin et al., "Prognostic factors of hepatocellular carcinoma patients treated by transarterial chemoembolization," *International Journal of Clinical and Experimental Pathology*, vol. 7, no. 11, pp. 1114–1123, 2014.

[21] J. J. Arends, H. P. Sleeboom, M. B. L. Leys et al., "A phase II study of raltitrexed and gemcitabine in patients with advanced pancreatic carcinoma," *British Journal of Cancer*, vol. 92, no. 3, pp. 445–448, 2005.

[22] J. M. Frakes, T. Strom, G. M. Springett et al., "Resected pancreatic cancer outcomes in the elderly," *Journal of Geriatric Oncology*, vol. 6, no. 2, pp. 127–132, 2015.

[23] M. Ernst, M. Najdovska, D. Grail et al., "STAT3 and STAT1 mediate IL-11-dependent and inflammation-associated gastric tumorigenesis in gp130 receptor mutant mice," *The Journal of Clinical Investigation*, vol. 118, no. 5, pp. 1727–1738, 2008.

[24] A. Arlt, A. Gehrz, S. Müerköster et al., "Role of NF-κB and Akt/PI3K in the resistance of pancreatic carcinoma cell lines against gemcitabine-induced cell death," *Oncogene*, vol. 22, no. 21, pp. 3243–3251, 2003.

Hindawi
BioMed Research International
Volume 2018, Article ID 1718046, 9 pages
https://doi.org/10.1155/2018/1718046

Research Article

Disease Sequences High-Accuracy Alignment Based on the Precision Medicine

ManZhi Li,[1] HaiXia Long,[2] HongTao Wang,[1] HaiYan Fu,[2] Dong Xu,[2] YouJian Shen,[1] YuHua Yao,[1] and Bo Liao[1]

[1]*School of Mathematics and Statistics, Hainan Normal University, Haikou, Hainan 571158, China*
[2]*School of Information Science Technology, Hainan Normal University, Haikou, Hainan 571158, China*

Correspondence should be addressed to HaiXia Long; 64169486@qq.com

Received 22 November 2017; Accepted 18 January 2018; Published 22 February 2018

Academic Editor: Tao Huang

High-accuracy alignment of sequences with disease information contributes to disease treatment and prevention. The results of multiple sequence alignment depend on the parameters of the objective function, including gap open penalties (GOP), gap extension penalties (GEP), and substitution matrix (SM). Firstly, the theory parameter formulas relating to GOP, GAP, and SM are inferred, combining unaligned sequence length, number, and identity. Secondly, we tested the rationality of the theory parameter formulas, with experiment on the ClustalW and MAFFT program. In addition, we obtained a group of MAFFT program parameters according to the formulas proposed. The results of all experiments show that the SPS (sum-of-pair score) obtained from theory parameters is better than the SPS obtained from the default parameters of ClustalW and MAFFT. In both theory and practice, our method to determine the parameters is feasible and efficient. These can provide high-accuracy alignment results for precision medicine.

1. Introduction

In 2015, US President Barack Obama stated his intention to fund a United States national "Precision Medicine Initiative" [1, 2]. A short-term goal of the Precision Medicine Initiative is to expand cancer genomics to develop better prevention and treatment methods. With the explosive growth of medical data, the complexity of disease, and the demand of personalized medicine, the research results of genome sequencing are changing the process of disease treatment. Multiple sequence alignment (MSA) is more and more important.

Multiple sequence alignment (MSA) has wide applications in sequence analysis, gene recognition, protein structure prediction, and reconstructing the phylogenetic tree [3]. Notredame [4] stated that the most modern programs for constructing MSA consist of two components: (1) an objective function to assess the quality of candidate alignment and (2) an optimization procedure for identifying the highest scoring alignment with respect to the chosen objective function. Currently, MSA has three main objective functions: (1) the sum-of-pairs score function (SPS), (2) the consensus function, and (3) the tree function. The SPS function is the most commonly used objective function, and its parameters include substitution matrix and gap opening penalties (GOP) and gap extending penalties (GEP).

The parameters of the objective function have generated many discussions on how to obtain optimal parameters. Thompson et al. [5] determined that substitution matrices vary at different alignment stages according to the divergence of sequences to be aligned. Residue-specific gap penalties and gap penalties in hydrophilic regions, which have been locally reduced, can cause new gaps to appear in potential loop regions rather than in a regular secondary structure. Reese and Pearson [6] discussed the relational formula between the PAM distance and PAM matrix as well as the gap penalty. Madhusudhan et al. [7] proposed the variable penalty formula according the structure of sequence based on dynamic programming. However, these formulas are not widely used. Gondro and Kinghorn [8] indicated that gap penalty parameters were determined by experience. At present, it is no theoretical framework to determine the optimum parameters. The current parameters pertaining

to the objective function in most literature are empirical values which are independently associated with the sequences [9]. BALiBASE is a database of manually refined multiple sequence alignments [10] and is usually used to test performance of MSA method [11].

Many open source online alignment tools are available that can align hundreds of thousands of sequences in hours. These include CLUSTAL Omega, T-COFFEE, and MAFFT, [5, 12–14] and often become the primary source of sequence alignment solution. However, these MSA tool results strongly depend on the gap penalty and substitution matrix. Different parameter combinations can obtain different MSA results. The majority of users use a single default parameter when applying these alignment tools, but the results are not the best. Moreover, an effective methodology has not yet been developed to directly determine an MSA optimal parameter, which means current online tools cannot guarantee the best solution. However, when compared with other MSA alignment tools, MAFFT has the advantage of simple input parameters and obtains better results than the other tools [12, 13]. This paper uses MAFFT as the basic experimental tool to verify the accuracy of the original formulas presented herein as they relate to the substitution matrix and the gap penalty.

2. Sum-of-Pairs (SP) Objective Function

The sum-of-pairs (SP) function is commonly used as an objective function for MSA and is derived as

$$\text{score} = \sum \text{Residue} - \sum \text{penalty}, \tag{1}$$

where the score is >0. When the score is higher, the accuracy of MSA is higher [15]. $\sum$ Residue > 0 represents the total score of amino acid residues in the alignment sequence. $\sum$ penalty is the total penalty score due to inserting gap and $\sum$ penalty $>$ 0.

$\sum$ Residue is calculated as

$$\sum \text{Residue} = \sum_{h=1}^{L} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \text{Cost}\left(S_i, S_j\right), \tag{2}$$

where S_{ih} is the h residue of the i sequence, L is the length of the aligned sequences, and k is the number of the sequences.

$$\text{Cost}\left(S_i, S_j\right) = \begin{cases} S_{aa} = \text{Score}\,(a, a) & \text{if } a = a \text{ (residues are matched)} \\ S_{ab} = \text{Score}\,(a, b) & \text{if } a \neq \text{“}-\text{”}, \; b \neq \text{“}-\text{”} \text{ (residues are mismatched)} \\ S_{a-} = \text{Score}\,(a, -) = 0 & \text{if } a \neq \text{“}-\text{”} \left(\text{residue and gap}\right). \end{cases} \tag{3}$$

Cost is computed by a substitution matrix. Currently, two main kinds of substitution matrices are available: PAM and BLOSUM. The BLOSUM series applies to this research. In substitution matrices, S_{aa} are different from each other. When the residues are mismatched, S_{ab} are also different from each other. But, in the process of simplifying the calculation, we need to use a precise and representative numerical value to represent the characteristics of the matrix. The average value can be a good characteristic representing a group of different data. Therefore, using the average value mean(S_{aa}) of S_{aa} represents the match of the matrix and using an average value mean(S_{ab}) of S_{ab} represents the mismatch of the matrix.

The calculation of $\sum$ penalty is divided into two categories: linear penalty and affine penalty. Linear penalty penalizes the same score for each gap. Affine penalty is commonly used because it is biologically meaningful [16–18]. The gap is divided into two types: gap open penalty (GOP) and gap extension penalty (GEP), so the affine penalty formula is given as

$$\sum \text{penalty} = N_{\text{GOP}} \cdot \text{GOP} + N_{\text{GEP}} \cdot \text{GEP}, \tag{4}$$

where N_{GOP} is the number of GOP, N_{GEP} is the number of GEP, and GOP > GEP.

3. The Theory Parameters Determination of SP Function for MSA

Symbol Description. The number of unaligned sequences is m. The length of the longest sequence is $\text{len}_{\max}$. The length of the shortest sequence is $\text{len}_{\min}$. The mean identity is iden. The number of amino acid residues matched is $\text{num}_{\text{match}} = (m(m-1)/2) \cdot \text{len}_{\min} \cdot \text{iden}$. After alignment, the number of gaps inserted into each sequence is num_{gap}.

Table 1 summarizes the ratio of the longest sequence and the number of gaps inserted into the sequence of each data set in BAliBASE 2.0 and BAliBASE 3.0. It shows that the number of gaps in the longest sequence is not more than 0.2 times the length of the longest sequence. That is, the number of gaps in each sequence is $\text{num}_{\text{gap}} \leq \text{int}\,(0.2 \cdot \text{len}_{\max}) + \text{len}_{\max} - \text{len}_{\min}$, and int is the rounding function. Figure 1 shows how the sequence length and the number of gaps num_{gap} are related.

Figure 1 is an example. If $\text{len}_{\text{align}} = 25$, $\text{len}_{\max} = 21$, and $\text{len}_{\min} = 7$, the number of gaps inserted into the longest sequence is $\text{num}_{\text{gap}} = \text{len}_{\text{align}} - \text{len}_{\max} = 25 - 21 = 4$, and the ratio between the sequence and gaps is ratio $= (\text{len}_{\text{align}} - \text{len}_{\max})/\text{len}_{\max} = 4/21 = 0.19$. The number of gaps in the sequence is $\text{num}_{\text{gap}} \leq \text{int}\,[0.2 \cdot \text{len}_{\max}]$. The number of gaps inserting the shortest sequence is $\text{num}_{\text{gap}} = \text{len}_{\text{align}} - \text{len}_{\min} = 25 - 7 = 18$, and the number of gaps in sequence is $\text{num}_{\text{gap}} \leq$

Table 1: Ratio of the longest sequence and the number of gaps inserted into the sequence.

	BALIBASE 2.0					BALIBASE 3.0	
Data set	Test 1	Test 2	Test 3	Ref 2	Ref 3	RV11	RV12
Mean (ratio)	0.0769	0.0764	0.0744	0.1439	0.1612	0.1938	0.0784

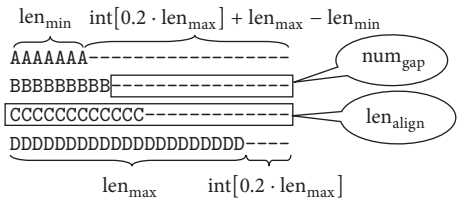


Figure 1: The relationship between the sequence length and the number of gaps.

$\text{int}[0.2 \cdot \text{len}_{\max}] + \text{len}_{\max} - \text{len}_{\min}$. The number of gaps in other sequences is $\text{num}_{\text{gap}} \le \text{int}[0.2 \cdot \text{len}_{\max}] + \text{len}_{\max} - \text{len}_{\min}$.

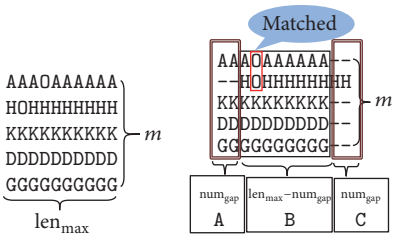
The following parameter formulas are inferred according to information obtained from Figure 2. Figure 2(a) has the best state unaligned sequence. Each sequence has the same length and no gaps. The longest length of any unaligned sequence is 10, so the number of gaps inserted can go up to 2. Figure 2(b) shows the worst alignment results (inserting maximum gap and minimum matching). If the score of Figure 2(b) is higher than the score of Figure 2(a), the parameters of the objective function meet all cases of alignment, because the situation in Figure 2 is the worst alignment.

3.1. Substitution Matrix Theory Formula. According to (1), the SP score of unaligned sequences is

$$\begin{aligned} \text{score}_{\text{begin}} &= \sum \text{Residue} - \sum \text{penalty} \\ &= \frac{m(m-1)}{2} \cdot \text{len}_{\max} \cdot S_{ab} \end{aligned} \tag{5}$$

and according to (1) and Figure 2(b), the following equations can be obtained:

$$\begin{aligned} &\text{score}_{\text{end}} = \text{score } A + \text{score } B + \text{score } C, \\ &\text{score } A = \alpha \cdot \frac{(m-1)(m-2)}{2} \cdot \text{num}_{\text{gap}} \cdot S_{ab}, \\ &\text{score } B \\ &\quad = \left[\frac{m(m-1)}{2} \left(\text{len}_{\max} - \text{num}_{\text{gap}} \right) - \text{num}_{\text{match}} \right] \\ &\qquad \cdot S_{ab} + \beta \cdot \text{num}_{\text{match}} \cdot S_{aa}, \\ &\text{score } C = \sum \text{penalty}. \end{aligned} \tag{6}$$



(a) The best state unaligned sequences (b) The worst alignment results

Figure 2: Unalignment and alignment.

So, the SP score of the aligned sequences is

$$\begin{aligned} \text{score}_{\text{end}} = &\left[\frac{m(m-1)}{2} \cdot \left(\text{len}_{\max} - \text{num}_{\text{gap}} \right) \right. \\ &\left. - \text{num}_{\text{match}} + \alpha \cdot \frac{(m-1)(m-2)}{2} \cdot \text{num}_{\text{gap}} \right] \cdot S_{ab} \\ &+ \beta \cdot \text{num}_{\text{match}} \cdot S_{aa} - \sum \text{penalty}. \end{aligned} \tag{7}$$

In theory, the alignment score must be greater than the unaligned sequence score,

$$\text{score}_{\text{begin}} \le \text{score}_{\text{end}}. \tag{8}$$

That is,

$$\begin{aligned} &\frac{m(m-1)}{2} \cdot \text{len}_{\max} \cdot S_{ab} \le \left[\frac{m(m-1)}{2} \right. \\ &\quad \cdot \left(\text{len}_{\max} - \text{num}_{\text{gap}} \right) - \text{num}_{\text{match}} + \alpha \\ &\quad \left. \cdot \frac{(m-1)(m-2)}{2} \cdot \text{num}_{\text{gap}} \right] \cdot S_{ab} + \beta \cdot \text{num}_{\text{match}} \\ &\quad \cdot S_{aa} - \sum \text{penalty}. \end{aligned} \tag{9}$$

Equation (9) can be simplified as

$$\begin{aligned} S_{aa} \ge &\left[\frac{(\alpha m - 2\alpha - m)(1-m)}{2\beta} \cdot \frac{\text{num}_{\text{gap}}}{\text{num}_{\text{match}}} + \frac{1}{\beta} \right] \\ &\cdot S_{ab}. \end{aligned} \tag{10}$$

The formula of the substitution matrix is shown in (10), which can be simplified as

$$\text{reference} \ge \text{calc}. \tag{11}$$

The rationality of the substitution matrix can be judged according to (11).

3.2. GOP and GEP Theory Formulas. Based on the affine penalty, num_{gap} is the number of gaps of each sequence; let us suppose that the number of gaps in each sequence is λ times as the number of GOP, so $N_{\text{GOP}} = m \cdot (1/\lambda) \cdot \text{num}_{\text{gap}}$ and

$N_{\text{GEP}} = m \cdot (1 - 1/\lambda) \cdot \text{num}_{\text{gap}}$. Because GOP > GEP, we accept that GOP = $n \cdot$ GEP, where λ, n is the positive integer, so

$$\begin{aligned} \sum \text{penalty} &= N_{\text{GOP}} \cdot \text{GOP} + N_{\text{GEP}} \cdot \text{GEP} \\ &= \frac{n + \lambda - 1}{n\lambda} \cdot m \cdot \text{num}_{\text{gap}} \cdot \text{GOP}. \end{aligned} \tag{12}$$

According to (12), (9) can be expressed as follows:

$$\begin{aligned} &\frac{(\alpha m - 2\alpha - m)(m-1)}{2} \cdot \text{num}_{\text{gap}} \cdot S_{ab} + \text{num}_{\text{match}} \\ &\cdot (\beta S_{aa} - S_{ab}) \geq \sum \text{penalty} \Longrightarrow \\ &\text{GOP} \leq \left[\frac{(\alpha m - 2\alpha - m)(m-1)}{2} \text{num}_{\text{gap}} S_{ab} \right. \\ &\left. + \text{num}_{\text{match}} (\beta S_{aa} - S_{ab}) \right] \\ &\cdot \frac{n\lambda}{m(n + \lambda - 1) \cdot \text{num}_{\text{gap}}}. \end{aligned} \tag{13}$$

Equation (13) is the upper limit of GOP and the lower limit is GOP > 0.

If the upper limit of GOP is multiplied by weight coefficient ω and $0 < \omega < 1$, the estimation formula of GOP is

$$\begin{aligned} \text{GOP} = \omega \cdot &\left[\frac{(\alpha m - 2\alpha - m)(m-1)}{2} \text{num}_{\text{gap}} S_{ab} \right. \\ &\left. + \text{num}_{\text{match}} (\beta S_{aa} - S_{ab}) \right] \\ &\cdot \frac{n\lambda}{m(n + \lambda - 1) \cdot \text{num}_{\text{gap}}}, \end{aligned} \tag{14}$$

where $\text{num}_{\text{match}} = (m(m-1)/2) \cdot \text{len}_{\min} \cdot \text{iden}$, $\text{num}_{\text{gap}} = \text{int}(0.2 \cdot \text{len}_{\max}) + \text{len}_{\max} - \text{len}_{\min}$, and int is a rounding function. $\text{len}_{\min}$ is the length of the shortest sequence in the unaligned sets, and iden is the mean identity of unaligned sets.

The estimation formula of GEP is

$$\text{GEP} = \frac{\text{GOP}}{n}. \tag{15}$$

The optimal value of each weight coefficients λ, n, ω, α, and β in (14) and (15) can be obtained through the following experiments.

4. Simulation and Results

In order to test the rationality of the parameter formulas and determine the optimal value of each weight coefficient, we designed the following experiments on the BAliBASE 2.0 and BAliBASE 3.0.

4.1. Experiment Setting. BAliBASE version 2.0 [10] is an improved version, extended from version 1 with 167 reference alignments to over 2100 sequences, which also features eight reference sets. Because all the reference alignments of BAliBASE are aligned by the manual, it often used to test algorithms [19–21]. Because our study is based on the global SP function, in this article, we used 113 reference alignments in References 1–3 as test objects. BAliBASE version 3.0 has the most widely used multiple alignment benchmark. The database contains 218 multiple protein sequence alignments, which have been divided into five reference sets. The first reference set includes equidistant sequences, whose identity is less than 20% (RV11) or between 20 and 40% (RV12) [22]. Other references have no similarity information. Because the formulas proposed in this paper need similarity of sequences, BAliBASE 2.0 and BAliBASE 3.0 (RV11 and RV12) were both used to establish data sets.

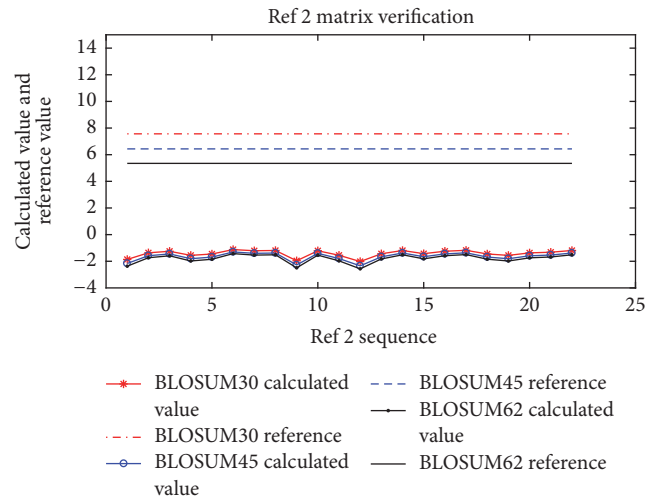


Figure 3: The results of the verification of substitution matrix (11).

SPS (sum-of-pair score) works as an objective function, which can determine score increases if sequences are correctly aligned. If the SPS is higher, the results of alignment are close to the reference alignment and can be even better than the reference alignment [20]. To test the rationality of presented formulas and to determine the optimal parameters combination of MSA tools, the most popular alignment program, MAFFT [16], is used in this research. The alignment results are obtained through the Perl programming language. The MAFFT program has some advantages: (1) the number of MAFFT program parameters is less and is easy to control, using only substitution matrices, GOP and GEP, (2) through Perl, the MAFFT program can batch align, and (3) alignment accuracy is for the most part better than CW, MUSCULE, and TCOFFEE.

In our experiment, $1 \leq \text{GOP} \leq 20, 0 \leq \text{GEP} \leq \text{GOP}/2$. The GOP step is 1, the GEP step is 0.2, and the substitution matrices are BLOSUM30, BLOSUM45, and BLOSUM62. For each group of sequences, through batch processing, the number of alignment results is 1,590 because there are 1,590 different combined parameter patterns.

4.2. Experiment Results

4.2.1. The Verification of Substitution Matrix Formula. This section shows how the rationality of the substitution matrix was established (see (11)). Figure 3 illustrates the calculated

Table 2: The number of sequences meeting the substitution matrix requirements (see (11)).

	Sequence number	Reference alignment number	BLOSUM30 qualified number (rate)	BLOSUM45 qualified number (rate)	BLOSUM62 qualified number (rate)
Reference 1	4-5	78	78 (100%)	78 (100%)	78 (100%)
Reference 2	14–19	22	22 (100%)	22 (100%)	22 (100%)
Reference 3	> 20	12	12 (100%)	12 (100%)	12 (100%)

Table 3: Determination of the value of $n, \lambda, \omega, \alpha, \beta$.

$n = 5, \lambda = 3, \alpha = 0.2, \beta = 0.9$	BLOSUM30		BLOSUM45		BLOSUM62	
ω	num	SPS	num	SPS	num	SPS
0.01	7	0.7586	11	0.7768	9	0.7697
0.02	9	0.761	10	0.7769	8	0.7692
0.03	10	0.7643	11	0.7795	10	0.7703
0.04	12	0.782	15	0.7886	12	0.7745
0.05	14	0.7843	**19**	**0.8003**	15	0.7846
0.06	14	0.7805	17	0.7924	17	0.7864
0.07	14	0.7767	16	0.7896	16	0.786
0.08	14	0.7728	16	0.784	14	0.7821
0.09	14	0.7668	15	0.7804	14	0.7777
0.1	14	0.764	15	0.78	15	0.7826
0.01	7	0.7586	11	0.7768	9	0.7697
0.02	9	0.7614	10	0.7769	9	0.77
0.03	10	0.7681	12	0.7818	12	0.7744
0.04	15	0.7874	15	0.7877	13	0.7858
0.05	13	0.7783	15	0.79	16	0.7918
0.06	13	0.7736	14	0.7845	14	0.7859
0.07	13	0.7732	14	0.7781	12	0.7819
0.08	13	0.7709	16	0.7846	12	0.7713
0.09	14	0.7779	15	0.7781	13	0.7751
0.1	14	0.7731	15	0.7795	14	0.7779

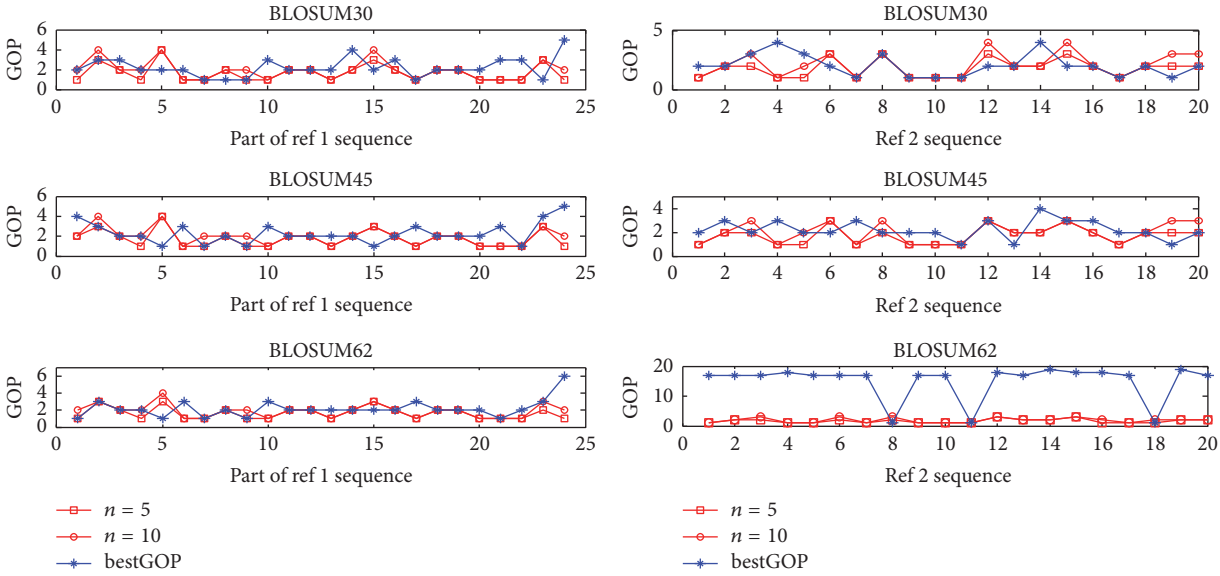
value and reference value of each of the three substitution matrices for Reference 2 (note: the other figures are similar to Figure 3). According to (11), when the reference value is greater than the reference value, the substitution matrix is rationality. It is shown that BLOSUM30, BLOSUM45, and BLOSUM62 meet the requirements of all sequences.

Table 2 lists the number of sequences meeting the substitution matrix sequence requirements (see (11)). It is shown that three BLOSUM substitution matrices meet all the sequences for References 1–3.

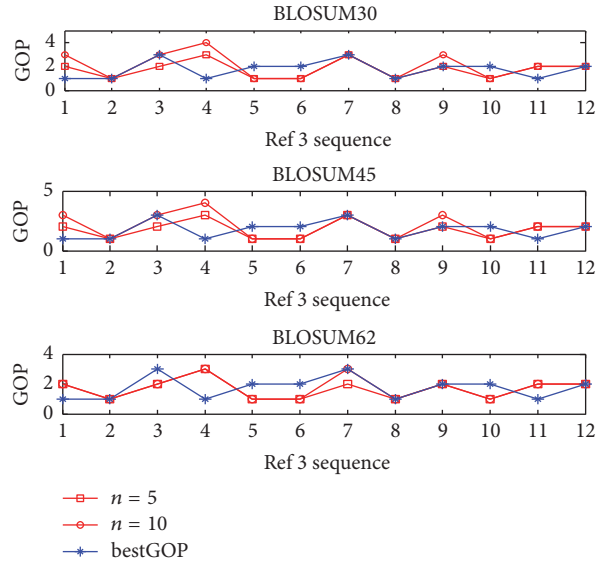
4.2.2. The Verification of Gap Penalty Formulas. Based on the SPS and MAFFT program (MAFFT-7.220-WIN64 version), we tested the rationality of (14) and (15). The optimum of GOP corresponded to the maximal SPS illustrated in Figure 4. From Figure 4, we can conclude the following: the GOP theory values inferred from (14) and (15) almost coincide with the optimal of GOP, so (14) can calculate the optimal value of GOP.

Table 3 statistics show the number of sequences in Reference 1 (Test 2), which meet the theory parameter requirements corresponding to SPS, which are greater than the default parameters corresponding to SPS. In Test 2, there are 24 sequences. Table 3 shows that when $\lambda = 3$, $\alpha = 0.2$, $\beta = 0.9$, and $n = 5$, the number of sequences is greater than $\lambda = 3, \alpha = 0.2, \beta = 0.9$, and $n = 10$. The best result is indicated in Blosum45, num 19, with an SPS of 0.8003 (in Table 3 set in bold face font). For Test 2 sequence sets, $\lambda = 3$, $n = 5$ is relatively rational and corresponds to $\omega = 0.05$. The other sequence sets can also obtain the value of n, λ, ω, α, and β, which are listed in Table 4.

4.2.3. Finding Optimal Value of Other Parameters in Derivation Formula. From the aforementioned experiments, we can determine the substitution matrix and n, λ, and ω in (14). The other parameters are related to the sequences where λ is the ratio of GOP and num_{gap}, and $\text{num}_{\text{gap}} = \text{int}(0.2 \cdot \text{len}_{\max}) + \text{len}_{\max} - \text{len}_{\min}$. The number of GOP is limited and it will



(a) The results of verification of GOP/GEP in Reference 1 sequences (b) The results of verification of GOP/GEP in Reference 2 sequences

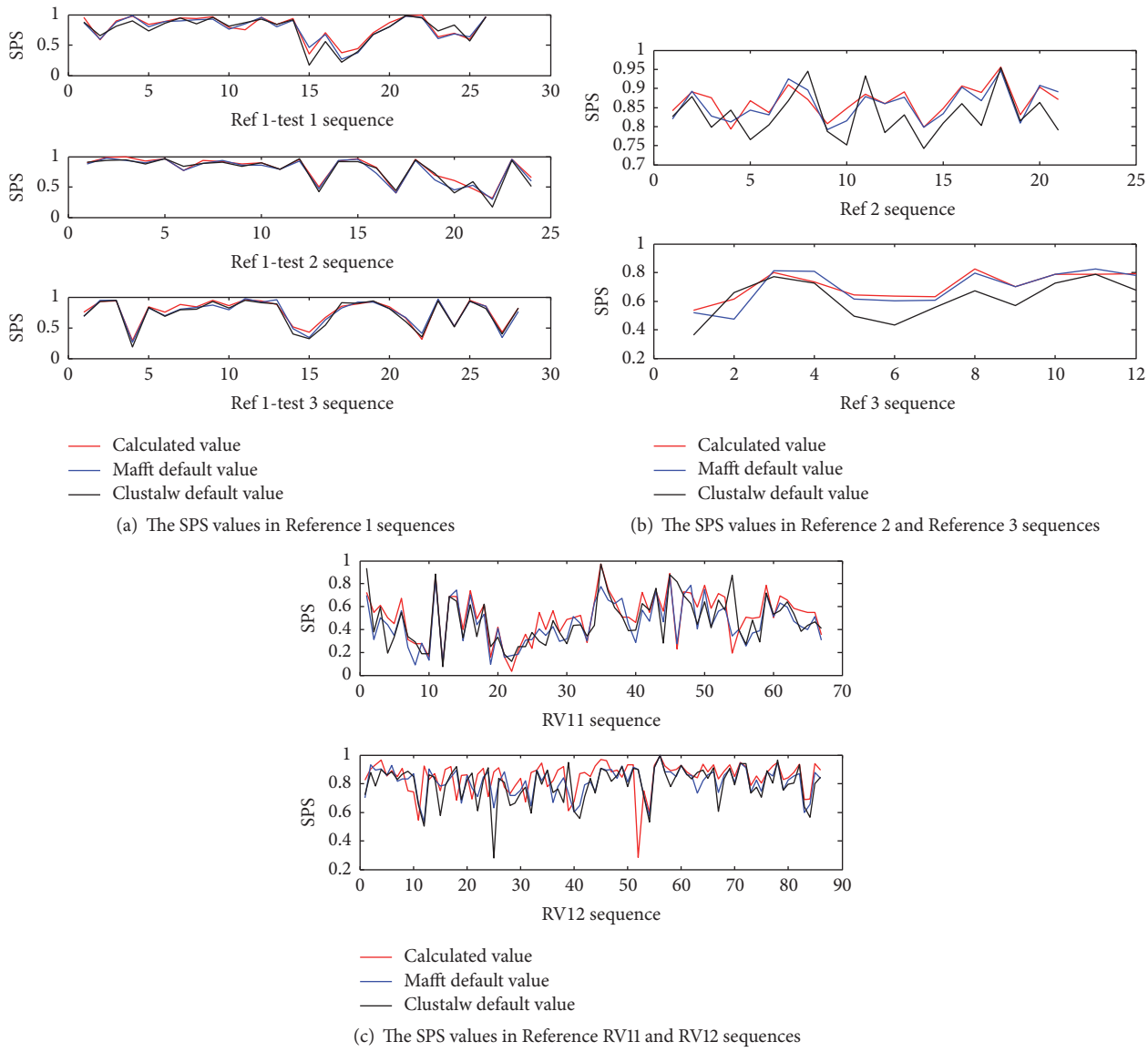


(c) The results of verification of GOP/GEP in Reference 3 sequences

Figure 4: The results of verification of GOP/GEP in (14) and (15).

Table 4: Optimal GOP/GEP/matrix.

Sequence set	Ref 1-test 1	Ref 1-test 2	Ref 1-test 3	Ref 2	Ref 3
Sequence row	4-5	4-5	4-5	14–19	>20
Sequence length (bp)	<100	100–300	>300	50–600	60–600
ω	0.03	0.05	0.08	0.02	0.02
n	5	5	10	10	10
Matrix	BLOSUM45	BLOSUM45	BLOSUM62	BLOSUM45	BLOSUM45
λ			3		
α			0.2		
β			0.9		



(a) The SPS values in Reference 1 sequences

(b) The SPS values in Reference 2 and Reference 3 sequences

(c) The SPS values in Reference RV11 and RV12 sequences

Figure 5: The SPS values are from MAFFT theory parameters, MAFFT default parameters, and CLUSTALW default parameter.

not increase too much, while the distribution of GEP is more concentrated. These parameters are more consistent with the biological characteristics of multiple sequence alignment.

Optimal parameters and the SPS value are listed in Table 4. The optimal value of weight coefficient in our proposed formula is located in Table 4. Using a weight coefficient, we can obtain the optimal of GOP, GEP, and MATRIX parameters. The number of sequences corresponding to SPS is also listed in Table 4.

Figure 5 shows that, for each SPS value sequence obtained from theory parameters, we inferred default parameters of MAFFT (MAFFT-7.220-WIN64 version) and CLUSTALW (CLUSTALW-2.1-WIN version). The SPS obtained by the MAFFT program are better than the CLUSTALW program on the default parameters. So we chose the MAFFT program as our test method. The SPS obtained by our theory parameters were better than the default parameters of MAFFT and CLUSTALW. Thus, the theory parameters we propose can optimize the results of MSA.

Table 5 shows the SPS mean values of References 1–3 sequences of BAliBASE 2.0 and RV11/RV12 of BAliBASE 3.0. The alignment sequences obtained from MAFFT default parameters, CLUSTALW default parameters, and MAFFT theory parameters are those proposed in this study. It is shown that SPS values obtained by MAFFT default parameters are better than SPS values obtained by CLUSTALW default parameters. The SPS values obtained using our theory parameters are the best. So, the theory parameters optimized the results of MSA.

5. Conclusions

This paper clearly shows that the parameters of MSA tools influence MSA results. These parameters not only include

Table 5: SPS mean value.

	BaliBASE 2.0					BaliBASE 3.0	
Data set	Ref 1 (test 1)	Ref 1 (test 2)	Ref 1 (test 3)	Ref 2	Ref 3	RV11	RV12
MAFFT default parameters	0.7749	0.7743	0.7460	0.8584	0.6938	0.4582	0.8142
CW default parameters	0.7614	0.7732	0.7340	0.8311	0.6189	0.4758	0.7966
MAFFT theory parameters	0.7918	0.8003	0.7652	0.8655	0.7073	0.5183	0.8449

substitution matrices, GOP, and GEP but also include the length, number, and identity of sequences. Our goal was to find a group of combined optimal parameters. Based on the SP function, we established a series of formulas which can determine the value of substitution, GOP, and GEP. In order to test the rationality of the formulas, our experiments were conducted in the MAFFT program base or in the BAliBASE 2.0 and BAliBASE 3.0 (RV11 and RV12) database. Moreover, we obtained the optimal value of the substitution matrices, GOP and GEP, and these values proved to be better than the default values of the MAFFT program. After the theory analysis and experimental analysis, we can conclude that the proposed method can effectively solve the MSA parameter problems and improve MSA accuracy, which can provide more accuracy information for precision medicine in disease analysis and prediction.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

ManZhi Li and HaiXia Long contributed equally to this work. ManZhi Li and HaiXia Long carried out the multiple sequence alignment parameters studies, participated in the experiments, and drafted the manuscript; these authors contributed equally to this work. HaiYan Fu and HongTao Wang participated in the design of the study and performed the statistical analysis. All authors read and approved the final manuscript.

Acknowledgments

This work was supported by the China Scholarship Council, the National Natural Science Foundation of China (no. 61762034, no. 71461008, no. 61663007, and no. 61163042), and the HaiNan Province Natural Science Foundation (no. 614235, no. 617122, and no. 20166222).

References

[1] T. P. Conrads and E. F. Petricoin, "The Obama Administration's Cancer Moonshot: A Call for Proteomics," *Clinical Cancer Research*, vol. 22, no. 18, pp. 4556–4558, 2016.

[2] F. S. Collins and H. Varmus, "A new initiative on precision medicine," *The New England Journal of Medicine*, vol. 372, no. 9, pp. 793–795, 2015.

[3] Q. Le, F. Sievers, and D. G. Higgins, "Protein multiple sequence alignment benchmarking through secondary structure prediction," *Bioinformatics*, vol. 33, no. 9, pp. 1331–1337, 2017.

[4] C. Notredame, "Recent progress in multiple sequence alignment: A survey," *Pharmacogenomics*, vol. 3, no. 1, pp. 131–144, 2002.

[5] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Research*, vol. 22, no. 22, pp. 4673–4680, 1994.

[6] J. T. Reese and W. R. Pearson, "Empirical determination of effective gap penalties for sequence comparison," *Bioinformatics*, vol. 18, no. 11, pp. 1500–1507, 2002.

[7] M. S. Madhusudhan, M. A. Marti-Renom, R. Sanchez, and A. Sali, "Variable gap penalty for protein sequence-structure alignment," *Protein Engineering, Design and Selection*, vol. 19, no. 3, pp. 129–133, 2006.

[8] C. Gondro and B. P. Kinghorn, "A simple genetic algorithm for multiple sequence alignment," *Genetics and Molecular Research*, vol. 6, no. 4, pp. 964–982, 2007.

[9] D. DeBlasio and J. Kececioglu, "Parameter advising for multiple sequence alignment," *BMC Bioinformatics*, vol. 16, no. Suppl 2, p. A3, 2015.

[10] J. D. Thompson, F. Plewniak, and O. Poch, "BAliBASE: A benchmark alignment database for the evaluation of multiple alignment programs," *Bioinformatics*, vol. 15, no. 1, pp. 87-88, 1999.

[11] R. C. Edgar, "Quality measures for protein alignment benchmarks," *Nucleic Acids Research*, vol. 38, no. 7, Article ID gkp1196, pp. 2145–2153, 2010.

[12] K. Katoh, K. Misawa, K. Kuma, and T. Miyata, "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform," *Nucleic Acids Research*, vol. 30, no. 14, pp. 3059–3066, 2002.

[13] K. Katoh and H. Toh, "Recent developments in the MAFFT multiple sequence alignment program," *Briefings in Bioinformatics*, vol. 9, no. 4, pp. 286–298, 2008.

[14] C. Notredame, D. G. Higgins, and J. Heringa, "T-coffee: a novel method for fast and accurate multiple sequence alignment," *Journal of Molecular Biology*, vol. 302, no. 1, pp. 205–217, 2000.

[15] K. Reinert, J. Stoye, and T. Will, "An iterative method for faster sum-of-pairs multiple sequence alignment," *Bioinformatics*, vol. 16, no. 9, pp. 808–814, 2000.

[16] O. Gotoh, "Multiple sequence alignment: Algorithms and applications," *Advances in Biophysics*, vol. 36, pp. 159–206, 1999.

[17] M. Kaya, A. Sarhan, and R. Alhajj, "Multiple sequence alignment with affine gap by using multi-objective genetic algorithm," *Computer Methods and Programs in Biomedicine*, vol. 114, no. 1, pp. 38–49, 2014.

[18] Q. Zou, X. Shan, and Y. Jiang, "A Novel Center Star Multiple Sequence Alignment Algorithm Based on Affine Gap Penalty and K-Band," *Physics Procedia*, vol. 33, pp. 322–327, 2012.

[19] A. Bahr, J. D. Thompson, J.-C. Thierry, and O. Poch, "BAliBASE (Benchmark Alignment dataBASE): Enhancements for repeats,

transmembrane sequences and circular permutations," *Nucleic Acids Research*, vol. 29, no. 1, pp. 323–326, 2001.

[20] F. M. Ortuño, O. Valenzuela, F. Rojas et al., "Optimizing multiple sequence alignments using a genetic algorithm based on three objectives: Structural information, non-gaps percentage and totally conserved columns," *Bioinformatics*, vol. 29, no. 17, pp. 2112–2121, 2013.

[21] F. Naznin, R. Sarker, and D. Essam, "Vertical decomposition with Genetic Algorithm for Multiple Sequence Alignment," *BMC Bioinformatics*, vol. 12, article no. 353, 2011.

[22] J. D. Thompson, P. Koehl, R. Ripp, and O. Poch, "BAliBASE 3.0: Latest developments of the multiple sequence alignment benchmark," *Proteins: Structure, Function, and Genetics*, vol. 61, no. 1, pp. 127–136, 2005.

Hindawi
BioMed Research International
Volume 2018, Article ID 6807059, 7 pages
https://doi.org/10.1155/2018/6807059

Research Article

Detecting Early Warning Signal of Influenza A Disease Using Sample-Specific Dynamical Network Biomarkers

Shanshan Zhu, Jie Gao, Tao Ding, Junhua Xu, and Min Wu

School of Science, Jiangnan University, Wuxi 214122, China

Correspondence should be addressed to Jie Gao; gaojie@jiangnan.edu.cn

Received 5 September 2017; Revised 30 November 2017; Accepted 25 December 2017; Published 31 January 2018

Academic Editor: Yudong Cai

Aims/Introduction. Evidences have shown that the deteriorated procession of disease is not a smooth change with time and conditions, in which a critical transition point denoted as predisease state drives the state from normal to disease. Considering individual differences, this paper provides a sample-specific method that constructs an index with individual-specific dynamical network biomarkers (DNB) which are defined as early warning index (EWI) for detecting predisease state of individual sample. Based on microarray data of influenza A disease, 144 genes are selected as DNB and the 7th time period is defined as predisease state. In addition, according to functional analysis of the discovered DNB, it is relevant with experience data, which can illustrate the effectiveness of our sample-specific method.

1. Introduction

A drastic change in the complex biological processes has been shown in recent studies, after which the system shifts rapidly from a stable state to another [1, 2]. This tipping point may be better known with respect to the earth climate system [3] and global finance [4] but now is applied in other areas gradually such as complex disease [5, 6]. In present studies, the disease progression is divided into three parts named normal state, predisease state (tipping point), and disease state, respectively [7]. Researches on predisease state of complex disease are only used to provide clinical patients with disease state the necessary information and are not to predict a patient in predisease state directly.

The earliest disease progression is identified by using a single molecular biomarker [8]. With further researches on disease progression, as early as 2008, Jin et al. applied the protein network to cardiovascular diseases, by identifying a group with high confidence of interacting proteins to form a network, which can be more accurate to divide into two groups of patients compared with a single molecular biomarker [9]. A more important role of network markers and a single molecular biomarker is to distinguish disease status, rather than to detect the critical state of the disease. Given this situation, Chen et al. proposed a theory of DNB to identify the critical state of the disease, which was based on model free, small sample, and high-throughput data. Three conditions to determine the DNB are put forward [10]. Generally, the studies of identifying predisease state are based on two types of data (high-throughput data and sequence data). Gao et al. extracted the larger mutation of influenza A virus proteins to form DNB based on sequence data with consecutive years, and according to the changes in DNB of each year, a warning index can be constructed to identify the outbreak year and before [11]. In addition, the development of high-throughput technology enables us to observe a large number of biomolecules by one time. Even if the number of patients' samples in the early state is small, it can also maintain each sampling point with high-throughput data on molecular level of high dimensionality [12–15].

Based on rapid advanced high-throughput technologies, we can obtain gene or protein expression at genome-wild scale with over thousands of measurements of long-term dynamics. Considering individual differences, our study is different from the method with multiple patient samples at each time period for detecting predisease state instead of proposing a sample-specific method [16–18]. In our study, the data sets which are divided into case group and control group were used to select differential expressed genes (DEGs) by

t-test. Genes in DEGs were clustered into 40 categories by using hierarchical clustering analysis. Then, 144 genes in a group which satisfied the three criteria of DNB identification proposed by Chen were selected as DNB. Therefore, based on individual-specific data, we can predict and identify whether a time period is in predisease state by observing the variation of EWI value combined with the three indicators.

2. Materials and Methods

2.1. Data Collection. The microarray gene expression data is downloaded from the National Center for Biotechnology Information's Gene Expression Omnibus (GEO) database (series accession number: GSE19392). The gene expression data set is generated by using Affymetrix HT Human Genome U133A (HT_ HG-U133A) Microarrays, which are obtained from an experiment of primary human bronchial epithelial cells that are infected with the wild-type PR8 influenza virus (A/PR/8/34). In our study, 10 out of 20 samples are defined as case group which are collected from primary human bronchial epithelial cells infected with the wild-type PR8 influenza virus after 0.25 h, 0.5 h, 1 h, 1.5 h, 2 h, 4 h, 6 h, 8 h, 12 h, and 18 h and the rest of the 20 samples are defined as control group which treated the same process but in the absence of virus. Moreover, 22277 probe sets are mapped to 13915 unique gene symbols involved in the influenza data set.

The Student t-test, which can evaluate the significance of genes with differential expression between case group and control group, is applied in the selection of DEGs. The p value figured out by t-test is directly used for the subsequent filtering analysis without multiple-testing correction. Only the genes with $p < 0.05$ are regarded as DEGs.

2.1.1. Dynamical System Model. Studies have shown that a biological process of the complex disease can be divided into three parts concretely. The state between normal and disease state is a tipping point which called predisease state. The system will change dramatically when the phase of disease approaches to the state. The following discrete-time state system of a living organism can be described by a nonlinear dynamical system equation:

$$\mathrm{A}\left(t+1\right) = f\left(A\left(t\right); p\right), \tag{1}$$

where $A(t) = (A_1(t), \ldots, A_n(t))'$ is an n-dimensional vector which represents observed values or molecule concentrations (e.g., gene expression or protein expression) at time point t ($t = 0, 1, \ldots$), for example, minutes, hours, or days. Parameter $P = (p_1, \ldots, p_s)$ indicates the slowly changing factors about genetic factors (e.g., SNP and CNV) and epigenetic factors (e.g., methylation and acetylation). Yet p is determined by its character and is not taken into consideration in this study because it is a unknown parameter with slower dynamics than $A(t)$. f is general nonlinear functions of $A(t)$.

2.1.2. Data Normalization. The observed values or molecule concentrations $A(t)$ can be classified into two groups, namely, the case group and the control group. They are denoted as $A_{\text{case}}(t)$ and $A_{\text{control}}(t)$, respectively:

$$\begin{aligned} A_{\text{case}}\left(t\right) &= \left(A_1^{\text{case}}\left(t\right), \ldots, A_n^{\text{case}}\left(t\right)\right)', \\ A_{\text{control}}\left(t\right) &= \left(A_1^{\text{control}}\left(t\right), \ldots, A_n^{\text{control}}\left(t\right)\right)'. \end{aligned} \tag{2}$$

Due to the existing large differences in the expression values of various genes or proteins, the data normalization manner as follows is adopted to analyze the data:

$$\widehat{a}_{nt} = \frac{A_n^{\text{case}}\left(t\right) - \text{mean}\left(\sum_t A_n^{\text{control}}\left(t\right)\right)}{\text{SD}\left(\sum_t A_n^{\text{control}}\left(t\right)\right)}, \tag{3}$$

where $A_n^{\text{case}}(t)$ is the nth gene or protein expression of case group and $\text{mean}(\sum_t A_n^{\text{control}}(t))$ and $\text{SD}(\sum_t A_n^{\text{control}}(t))$ are the mean and standard deviation of nth gene or protein expression at all time points in control group and the control group, respectively. Then a $n \times t$ normalization matrix is obtained:

$$\widetilde{A} = \begin{bmatrix} \widetilde{a}_{11} & \widetilde{a}_{12} & \cdots & \widetilde{a}_{1t} \\ \widetilde{a}_{21} & \widetilde{z}_{22} & \cdots & \widetilde{a}_{2t} \\ \vdots & \vdots & \vdots & \vdots \\ \widetilde{a}_{n1} & \widetilde{a}_{n2} & \cdots & \widetilde{a}_{nt} \end{bmatrix}, \tag{4}$$

where $\widetilde{a}_{nt}$ represents the normalization data of the nth gene or protein at time point t.

2.1.3. Sample-Specific Dynamical Network Biomarkers Selection. To further filtrate DNB, DEGs by t-test are isolated from normalization data $\widetilde{A}$ and denoted as $\widetilde{A}_1$, while the rest are denoted as $\widetilde{A}_2$. It is assumed that $\widetilde{A}_1$ and $\widetilde{A}_2$ are $r \times t$ matrix and $(n-r) \times t$ matrix, respectively. And $\widetilde{A}_1$ is clustered into 40 categories by using hierarchical clustering analysis. The Euclidean distance is applied to calculate the distance within genes or proteins of $\widetilde{A}_1$. The optimal group of genes or proteins is selected as DNB according to the following three criteria of DNB identification proposed by Chen:

(i) The average standard deviation (SD) of molecule concentration ($\widetilde{A}_i$) in this group is significantly higher comparing to others.

(ii) The average Pearson correlation coefficient (PCC) in absolute value of molecule concentrations ($\widetilde{A}_i$) in this group is relatively higher than the PCC between other molecules.

(iii) The average Pearson correlation coefficient in absolute value between molecule concentrations inside this group ($\widetilde{A}_i$) and anyone outside this group ($\widetilde{A}_j$) (OPCC) is much lower.

2.1.4. Construct Sample-Specific Early Warning Index. The optimal group containing q genes or proteins is separated from $\widetilde{A}_1$, which is marked as $\widetilde{A}_{\text{DNB}}$. Additionally, the rest of the groups of $\widetilde{A}$ are assigned to $\widetilde{A}_{\text{other}}$. There is a key point called predisease state during the development of the disease, in the figure of dynamic state (Figure 1(c)) originally

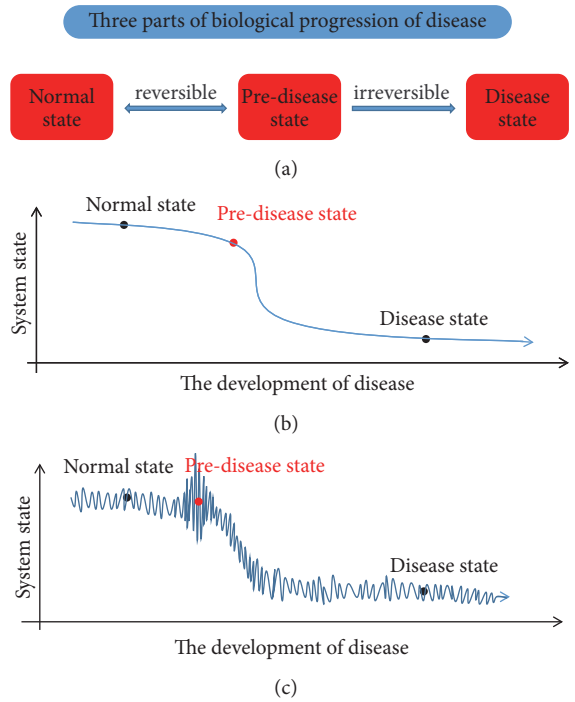


FIGURE 1: Three parts of biological progression of disease of a living organism. (a) The disease progression consists of three states including normal state, predisease state, and disease state, respectively. As shown in the picture, the process from normal to predisease state is reversible, whereas the process from predisease to disease state is irreversible. (b) The static variation displays the development progression of disease and the average value of molecule concentrations (e.g., gene or protein expression) at each state. (c) The dynamic variation shows the development progression of disease and the dynamic value of molecule concentrations (e.g., gene or protein expression) at each state.

proposed by Chen et al. [10]. The change of DNB before and after predisease state is relatively stable and smooth, whereas it turns into abrupt and drastic at predisease state. After identifying the DNB, the early warning index of each time point t can be constructed by three criteria:

(i) The average coefficient variation (CV) of molecule concentrations at different time points is the value of fluctuation. The CV value approaching predisease state is higher than that of other time point.

(ii) The average value of absolute difference (DIF) in molecule concentrations inside DNB approaching predisease state drastically decreases compared with the values at other time points.

(iii) The average value of absolute difference between molecule concentrations inside DNB and any other outside DNB (ODIF) approaching predisease state is relatively higher than others.

Hence, the EWI_t of all time points can be constructed as

$$\mathrm{EWI}_t = \frac{\mathrm{CV}_t \times \mathrm{DIF}_t}{\mathrm{ODIF}_t}, \tag{5}$$

where

$$\mathrm{CV}_t = \frac{\mathrm{SD}\left(\widetilde{A}_{\mathrm{DNB}}(t)\right)}{\mathrm{mean}\left(\widetilde{A}_{\mathrm{DNB}}(t)\right)} \tag{6}$$

$$\mathrm{DIF}_t = \frac{\sum_{i_1,i_2} \left|\widetilde{a}_{i_1 t} - \widetilde{a}_{i_2 t}\right|}{i_1 \times i_2} \tag{7}$$

$$(i_1, i_2 = 1, 2, \ldots, \text{ the number of DNB})$$

$$\mathrm{ODIF}_t = \frac{\sum_{i,j} \left|\widetilde{a}_{it} - \widetilde{a}_{jt}\right|}{i \times j} \tag{8}$$

$$(j = 1, 2, \ldots, \text{ the number of genes or proteins outside DNB}).$$

In the light of the characteristic in predisease state, a time point with the largest value can be considered as the predisease state. After the point, the disease progression of a living organism shifts rapidly from normal to disease state.

3. Result

All without-correct-corresponding gene symbols are screened out, and probe of the same genes is combined by the averaging method. There are 13915 genes left. Based on the 13915 genes, Student's t-test is applied to calculate the p value of each gene by comparing its expression profile between case groups and control groups. We identify 264 genes with $p < 0.05$ as DEGs. Next, 264 genes are classified by hierarchical clustering analysis into 40 categories. Analyzing all clusters or groups, a group of 144 genes is identified as DNB, which satisfies the three criteria of DNB identification. Among them, the values of average SD and average PCC in this group are 1.2585797 and 0.3047569, which is higher than others (e.g., 0.8802955 and 0.2940955), and the average OPCC is relatively high.

To further clarify the early warning index for influenza A disease with 10 time points, Figure 2 demonstrates variation of four indicators in detail. As shown in (a), the curve of the

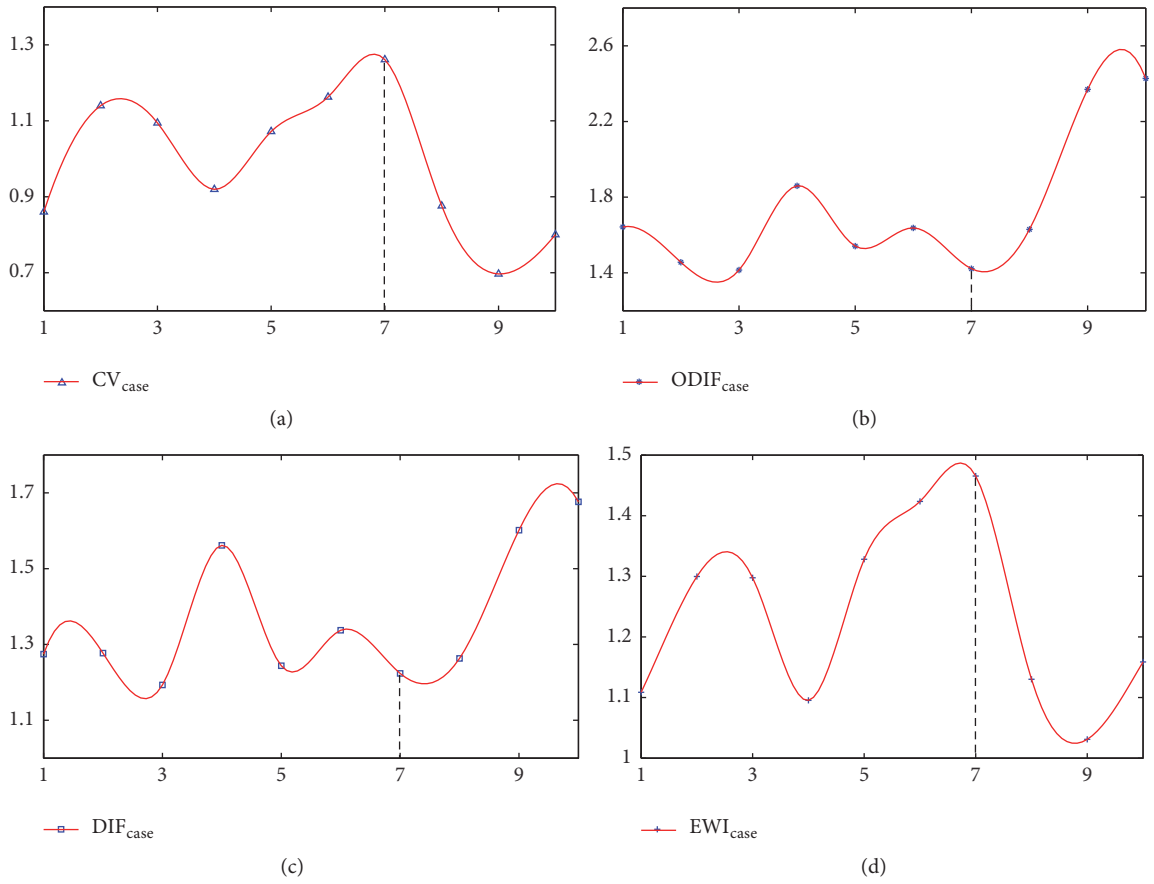


Figure 2: The early warning index of influenza disease. In all the figures, the abscissa represents the time point t. (a) The average coefficient variation (CV) of genes expression in DNB at 10 time points. (b) The average difference in absolute value between genes expression inside DNB and any other outside DNB (ODIF) at 10 time points. (c) The average difference (DIF) in absolute value of genes expression in DNB at 10 time points. (d) The early warning index (EWI) of the case set of high-throughput experimental data for influenza A disease.

CV value of DNB strongly fluctuates with time and the value at 7th time point (6 h) reaches the maximum value, which indicates that the genes in DNB change drastically when approaching 7th time point (6 h). And the value of DIF at 6 h shows the relatively lower value, which indicates that the trend of genes expression in DNB is similarly approaching 7th time point (6 h). Although the change in ODIF is not obvious, the early warning index at 7th time point (6 h) reaches the maximum value. Thus, the most prominent physiological effects occur approaching 7th time point (6 h). Meanwhile, gene expression changes in HBECs in response to wild-type influenza (PR8) show a strong float after 7th time point (6 h) [19].

In order to analyze biological functional of the obtained DNB, A bioinformatics database DAVID [20] (https://david.abcc.ncifcrf.gov/) with Gene Ontology (GO) analysis and KEGG Pathway analysis is mentioned. Some enriched GO functions based on identified genes in DNB are listed in Table 1. Gene Ontology can be divided into three parts: molecular function, biological process, and cellular composition. The analysis of genes reveals that the DNB selected by the sample-specific method is particularly related to influenza disease, which confirms the validation of our theory about the increasing index approaching predisease state. The enriched GO functions underlying the identified DNBs are particularly related to immune systems that are activated to protect against influenza A virus and inordinate dysfunctions associated with the performance in the viral life cycle. In the DNBs, DCAF1 and ADGRG3, which play crucial roles in cell differentiation, and ARVCF, COL19A1, and OLR1, which are associated with cell adhesion, regulate the expression of cell adhesion molecules [21]. Further, some genes in the basic cellular processes are expressed in a disorderly manner, for example, IFNA10, which is associated with the regulation of cell death and abnormal reaction in transcription and translation. Moreover, Some of them are involved in the related triglyceride metabolic process, especially for APOC3.

According to KEGG Pathway enrichment analysis, the results show that genes in DNB of influenza A disease are closely relevant to immune system and inflammation, such as cytokine-cytokine receptor interaction, PPAR signaling pathway, and Jak-STAT signaling pathway in Table 2. As key genes in cytokine-cytokine receptor interaction, CXCR6, IFNA10, IL21R, and TGFB3 in DNB participate in immune response and immune regulation, regulate innate immune and adapt

Table 1: Functional enrichment of GO for part of genes of identified DNB.

GO term	Description	DNB	p value	Corrected p value
GO:0007155	Cell adhesion	ARVCF, COL19A1, OLR1, CD300A, PCDHA10, CX3CL1, OMG, SIGLEC9	0.03844	0.85044
GO:0045766	Positive regulation of angiogenesis	ADM2, NOS3, CX3CL1, ALOX12	0.04448	0.86098
GO:0005886	Plasma membrane	SCN3B, GRIK2, TGFB3, SLC7A9, DMPK, APOA1,...	$9.55E-7$	$1.56E-4$
GO:0005615	Extracellular space	LPL, CRISP3, LUM, AFM, IFNA10, TGFB3, CX3CL, BMP15, APOA1, APOL1, IL18BP,...	$1.23E-4$	0.00400
GO:0005102	Receptor binding	HAO1, LPL, TENM2, HAO2, CX3CL1, LTB	0.03689	0.97736
GO:0010181	FMN binding	HAO1, HAO2, NOS3	0.00418	0.67592

Table 2: Functional enrichment of KEGG pathways for part of genes of identified DNB.

Pathway term		DNB	p value
hsa04060	Cytokine-cytokine receptor interaction	CXCR6; IFNA10; IL21R; TGFB3; LTB; EDA2R; CX3CL1; IL13RA1; IL3RA	$3.24E-4$
hsa03320	PPAR signaling pathway	LPL; APOA1; OLR1; APOC3	$1.45E-2$

immune response, stimulate hematopoietic function, stimulate cell activation, proliferation, and differentiation, and induce apoptosis. The pathway of cytokine-cytokine receptor interaction is the same with expressed data.

Moreover, the genes in PPAR signaling pathway like LPL, APOA1, OLR1, and APOC3 play a significant role in inhibiting inflammation, regulating cell apoptosis and immune system. And the genes in DNB which are marked red are placed the critical positions in cytokine-cytokine receptor interaction and PPAR signaling pathway. As shown in Figure 3. JAK-STAT signaling pathway is a signal transduction pathway stimulated by cytokines in recent years [22], which includes IFNA10, IL21R, IL13RA1, and IL3RA in DNB, involved in cell proliferation, differentiation, apoptosis, and immune regulation, and many other important biological processes.

To further demonstrate the effectiveness of our method, we analyze symptoms of patients and their complications. Patients develop symptoms of illness of upper respiratory tract infection. However, they are also accompanied by the occurrence of pulmonary complications, and renal failure [23]. Moreover, 18 out of 144 genes are validated with significantly close relation with influenza A disease. The CX3CL1 involves both acute and chronic inflammations, which is characterized by major perturbations of the immune homeostasis [24]. Especially surfactant proteins SFTPB plays a key role in alveolar stability [25], Which is associated with influenza A disease. And this gene encodes the pulmonary-associated surfactant protein B. The surfactant is secreted by the alveolar cells of the lung and maintains the stability of pulmonary tissue by reducing the surface tension of fluids that coat the lung.

4. Discussion

To detect the early warning signal of influenza A disease using a small number of samples of high-throughput data, we propose an early warning index serving as a leading indicator to predict the critical transition based on the concept of dynamical network biomarkers proposed by Chen, which drives the disease progression from normal state to disease state. Compared to the general biomarkers [26], dynamical network biomarkers are more suitable for characterizing the transfer of system status. In our study, We first select the DEGs by t-test between case groups and control groups. Then, a new type of normalization data is constructed by the formula defined in this study for the sake of analysis of the next step. Different from the previous methods, our work regards the gene expression with time of each gene as a vector for hierarchical clustering analysis. And the Euclidean distance is applied to calculate the distance within genes in DEGs. A group, which satisfies three criteria of DNB identification, is identified as DNB. Further, the values of CV, DIF, and ODIF are calculated to construct an index for detecting predisease state of individual sample. The index EWI is applied in early diagnosis with the microarray data of influenza A disease, which demonstrates fluctuated values with time. Although the ODIF value approaching predisease state is not completely obvious, the expression value of the other three indicators is significantly relevant with our theory. In addition, everyone with the same disease has different DNB due to different driving factors. We will focus on this important future topic and continue to refine the algorithm in later research.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study is supported by Major Research Plan of National Natural Science Foundation of China under Grant no. 91730301; National Natural Science Foundation of China

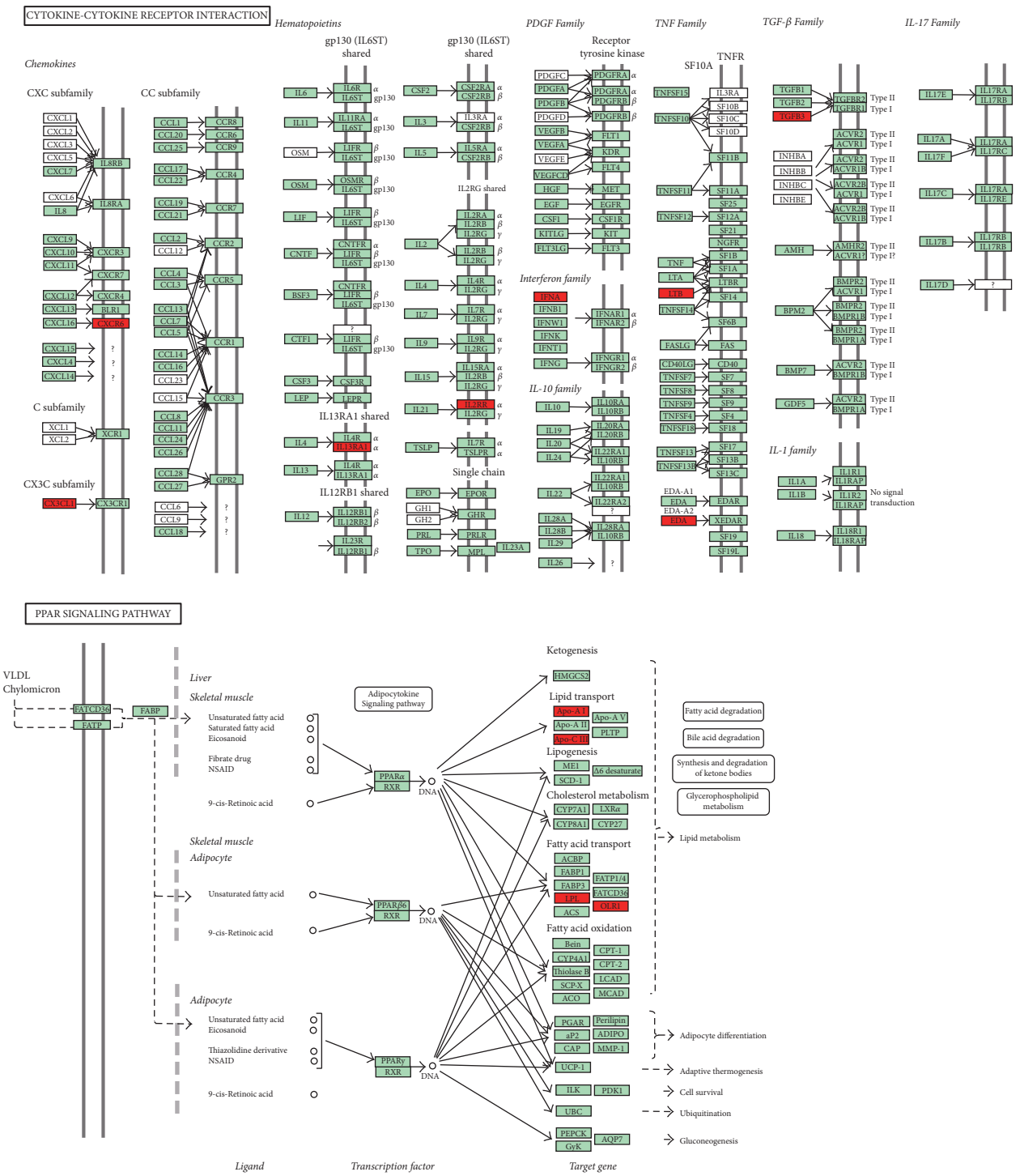


Figure 3: Key biological pathways with DNB genes in cytokine-cytokine receptor interaction and PPAR signaling pathway.

under Grant no. 11271163; the Foundation of the Innovation Project of Jiangsu Province under Grant no. KYCX17_1478.

References

[1] W. Joost Lesterhuis, A. Bosco, M. J. Millward, M. Small, A. K. Nowak, and R. A. Lake, "Dynamic versus static biomarkers in cancer immune checkpoint blockade: Unravelling complexity," *Nature Reviews Drug Discovery*, vol. 16, no. 4, pp. 264–272, 2017.

[2] M. A. Dahlem, J. Kurths, M. D. Ferrari, K. Aihara, M. Scheffer, and A. May, "Understanding migraine using dynamic network biomarkers," *Cephalalgia*, vol. 35, no. 7, pp. 627–630, 2015.

[3] A. Nikolaou, P. A. Gutiérrez, A. Durán, I. Dicaire, F. Fernández-Navarro, and C. Hervás-Martínez, "Detection of early warning

signals in paleoclimate data using a genetic time series segmentation algorithm," *Climate Dynamics*, vol. 44, no. 7-8, pp. 1919–1933, 2015.

[4] D. Karahoca, A. Karahoca, and Ö. Yavuz, "An early warning system approach for the identification of currency crises with data mining techniques," *Neural Computing and Applications*, vol. 23, no. 7-8, pp. 2471–2479, 2013.

[5] R. Liu, M. Li, Z.-P. Liu, J. Wu, L. Chen, and K. Aihara, "Identifying critical transitions and their leading biomolecular networks in complex diseases," *Scientific Reports*, vol. 2, article no. 813, 2012.

[6] P. Chen, R. Liu, Y. Li, and L. Chen, "Detecting critical state before phase transition of complex biological systems by hidden Markov model," *Bioinformatics*, vol. 32, no. 14, pp. 2143–2150, 2016.

[7] A. Achiron, I. Grotto, R. Balicer, D. Magalashvili, A. Feldman, and M. Gurevich, "Microarray analysis identifies altered regulation of nuclear receptor family members in the pre-disease state of multiple sclerosis," *Neurobiology of Disease*, vol. 38, no. 2, pp. 201–209, 2010.

[8] J. S. Kinney, T. Morelli, T. Braun et al., "Saliva/pathogen biomarker signatures and periodontal disease progression," *Journal of Dental Research*, vol. 90, no. 6, pp. 752–758, 2011.

[9] G. Jin, X. Zhou, H. Wang et al., "The knowledge-integrated network biomarkers discovery for major adverse cardiac events," *Journal of Proteome Research*, vol. 7, no. 9, pp. 4013–4021, 2008.

[10] L. Chen, R. Liu, Z.-P. Liu, M. Li, and K. Aihara, "Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers," *Scientific Reports*, vol. 2, article 342, 2012.

[11] J. Gao, L. Zhang, and P. Jin, "Influenza pandemic early warning research on HA/NA protein sequences," *Current Bioinformatics*, vol. 9, no. 3, pp. 228–233, 2014.

[12] T. Zeng, S.-Y. Sun, Y. Wang, H. Zhu, and L. Chen, "Network biomarkers reveal dysfunctional gene regulations during disease progression," *FEBS Journal*, vol. 280, no. 22, pp. 5682–5695, 2013.

[13] R. Liu, X. Wang, K. Aihara, and L. Chen, "Early diagnosis of complex diseases by molecular biomarkers, network biomarkers, and dynamical network biomarkers," *Medicinal Research Reviews*, vol. 34, pp. 455–478, 2014.

[14] M. Li, T. Zeng, R. Liu, and L. Chen, "Detecting tissue-specific early warning signals for complex diseases based on dynamical network biomarkers: Study of type 2 diabetes by cross-tissue analysis," *Briefings in Bioinformatics*, vol. 15, no. 2, pp. 229–243, 2014.

[15] A. D. Torshizi and L. Petzold, "Sparse Pathway-Induced Dynamic Network Biomarker Discovery for Early Warning Signal Detection in Complex Diseases," *IEEE/ACM Transactions on Computational Biology & Bioinformatics*, pp. 1–8, 2017.

[16] X. Liu, Y. Wang, H. Ji, K. Aihara, and L. Chen, "Personalized characterization of diseases using sample-specific networks," *Nucleic Acids Research*, vol. 44, no. 22, article no. e164, 2016.

[17] X. Liu, X. Chang, R. Liu, X. Yu, L. Chen, and K. Aihara, "Quantifying critical states of complex diseases using single-sample dynamic network biomarkers," *PLoS Computational Biology*, vol. 13, no. 7, Article ID e1005633, 2017.

[18] R. Liu, X. Yu, X. Liu, D. Xu, K. Aihara, and L. Chen, "Identifying critical transitions of complex diseases based on a single sample," *Bioinformatics*, vol. 30, no. 11, pp. 1579–1586, 2014.

[19] S. D. Shapira, I. Gat-Viks, B. O. V. Shum et al., "A physical and regulatory map of host-influenza interactions reveals pathways in H1N1 infection," *Cell*, vol. 139, no. 7, pp. 1255–1267, 2009.

[20] B. T. Sherman, D. W. Huang, Q. Tan et al., "DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis," *BMC Bioinformatics*, vol. 8, article 426, 2007.

[21] D. Zhang, N. Tang, Y. Liu, and E.-H. Wang, "ARVCF expression is significantly correlated with the malignant phenotype of non-small cell lung cancer," *Molecular Carcinogenesis*, vol. 54, no. 1, pp. E185–E191, 2015.

[22] S.-F. Hsu, W.-C. Su, K.-S. Jeng, and M. M. C. Lai, "A host susceptibility gene, DR1, facilitates influenza a virus replication by suppressing host innate immunity and enhancing viral RNA replication," *Journal of Virology*, vol. 89, no. 7, pp. 3671–3682, 2015.

[23] A. Antonopoulou, F. Baziaka, T. Tsaganos et al., "Role of tumor necrosis factor gene single nucleotide polymorphisms in the natural course of 2009 influenza A H1N1 virus infection," *International Journal of Infectious Diseases*, vol. 16, no. 3, pp. e204–e208, 2012.

[24] W. Liu, L. Jiang, C. Bian et al., "Role of CX3CL1 in Diseases," *Archivum Immunologiae et Therapia Experimentalis*, vol. 64, no. 5, pp. 371–383, 2016.

[25] K. K. W. To, J. Zhou, Y.-Q. Song et al., "Surfactant protein B gene polymorphism is associated with severe influenza," *CHEST*, vol. 145, no. 6, pp. 1237–1243, 2014.

[26] Z.-P. Liu, "Identifying network-based biomarkers of complex diseases from high-throughput data," *Biomarkers in Medicine*, vol. 10, no. 6, pp. 633–650, 2016.

Hindawi
BioMed Research International
Volume 2018, Article ID 5920608, 8 pages
https://doi.org/10.1155/2018/5920608

Research Article

Prognostic Value of Immunoscore and PD-L1 Expression in Metastatic Colorectal Cancer Patients with Different RAS Status after Palliative Operation

Ruiqi Liu, Ke Peng, Yiyi Yu, Li Liang, Xiaojing Xu, Wei Li, Shan Yu, and Tianshu Liu

Department of Oncology, Zhongshan Hospital, Fudan University, Shanghai 200032, China

Correspondence should be addressed to Tianshu Liu; liutianshu1969@126.com

Received 28 November 2017; Accepted 27 December 2017; Published 31 January 2018

Academic Editor: Jialiang Yang

Colorectal cancer (CRC) is the fifth leading cause of cancer death and the fifth most commonly diagnosed cancer in China. Approximately, 25% of CRC was in the advanced stage as diagnosed, and 40% of patients with CRC progress to metastatic colorectal cancer (mCRC). RAS mutation status is now routinely used to select their therapy. But it is still a question whether RAS mutation status is a prognostic marker. In our study, we detected RAS mutation, immunoscore (IS), and PD-L1 expression in 60 Chinese mCRC patients who received palliative operation. The Kaplan-Meier survival analysis showed that the overall survival (OS) in patients with RAS wild type was better than those with RAS mutated type. Moreover, in multivariate analysis, RAS mutation and PD-L1 expression were demonstrated to be the independent negative prognostic factors for OS ($P = 0.044$, HR: 0.258, and 95% CI: 0.069–0.967; $P = 0.048$, HR: 0.276, and 95% CI: 0.077–0,988). All results suggested that, combined with IS, PD-L1 expression and RAS status may be the prognostic indicators for mCRC patients with palliative operation.

1. Introduction

The World Health Organization (WHO) showed nearly half of colorectal cancer (CRC) cases are detected in Asia, mostly in China. CRC was the fifth most commonly diagnosed cancer in China [1], with more than 0.3 million new cases and 191000 deaths occurring [2]. In the last few years, the mortality of CRC was declining in United States but rapidly growing in China, which is the fifth leading cause of cancer death. Furthermore, approximately 25% of CRC was in the advanced stage as diagnosed and more than 40% of patients with CRC progress to metastatic colorectal cancer (mCRC) [3].

The RAS protooncogenes encode a family of highly homologous proteins, including HRAS, KRAS, and NRAS. They are involved in RAS/RAF/MEK/ERK signal pathway, which regulates the growth and survival properties of cells [4]. For mCRC patients, RAS mutation is usually used as an important predictive factor for the clinical response of anti-EGFR treatment. Recent studies have demonstrated that BRAF mutations are related to poor prognosis of mCRC [5–7]. However, we could not draw a firm conclusion about the correlation between the RAS mutation and the prognosis in mCRC patients with palliative operation.

Tumor-infiltrating immune cells, which play a role in recognition and elimination of tumor cell, have been reported to promote immune evasion and metastasis in CRC [8, 9]. Recently, several studies have demonstrated that immunoscore (IS), based on the density of CD8+ and CD3+ tumor-infiltrating lymphocytes in the invasive margin and the core of tumor, is vastly thought to be superior to the current tumor-node-metastases (TNM) staging system [10, 11]. However, the evidence is limited for mCRC.

Programmed cell death-ligand 1 (PD-L1) has been reported to function in the immunoregulatory system during certain conditions, including autoimmune disease, allograft rejection, pregnancy, and cancer [12]. Several studies suggested that PD-L1 expression in lymphocyte cells and in tumor cells of CRC is related to a high density of tumor-infiltrating immune cells [13, 14]. Hence, expression levels of PD-L1 were inversely correlated to T-cell densities in CRC tissue. However, the complex interrelationship between prognostic of mCRC and PD-L1 expression is still unknown.

Although most studies have demonstrated that BRAF mutations are related to poor prognosis of mCRC, we could not draw a firm conclusion about the correlation between the RAS mutation and the prognosis in mCRC patients. The objectives of this study were to confirm the prognostic value of the immunoscore of CD3+CD8 and the PD-L1 expression in mCRC with or without RAS mutation.

2. Materials and Method

2.1. Patients. This retrospective study included 60 mCRC patients with palliative operation at diagnosis between December 2013 and March 2016. Available variables included the following: sex, age of diagnosis, tumor location, RAS mutation type, histological type, vascular and perineural invasion, and metastatic sites. All patients were followed up until their deaths, or their last follow-up, or March 31, 2017. We defined the overall survival (OS) as the time from the date of primary treatment to the date of the last follow-up.

2.2. Immunohistochemistry and Image Analysis of Tumor-Infiltrating Immune Cell. The presence of tumor-infiltrating immune cells was confirmed by immunohistochemistry using antibodies for CD3 (ZA-0503), CD8 (ZA-0508), and PD-L1 (ab205921). Immunostaining for CD3 and CD8 and PD-L1 was performed using a Bond polymer kit (Leica Microsystems) and Leica BONDMAX autostainer (Leica Microsystems). All immunostained slides were scanned on an Aperio ScanScope® CS instrument (Aperio Technologies, Inc., Vista, CA, USA). The immunomarker-positive tumor-infiltrating immune cells were quantified by computerized image analysis system, ImageScope™ (Aperio Technologies). CD3+, CD8+, and PD-L1+ lymphocytes were counted using the Nuclear v9 algorithm. The density of immune infiltrates was obtained from the entire area of the tissue core.

2.3. Determination of Scoring System. Immunoscore (IS) was performed as described before [15]. Briefly, immunomarker-positive tumor-infiltrating immune cells were quantified by computerized image analysis system, ImageScope (Aperio Technologies). CD3+ and CD8+ lymphocytes were counted using the Nuclear v9 algorithm. We used the same cut-off values as Kwak et al. described. IS was defined as a quantification system based on the combination of two markers (CD3 and CD8) in two regions—the core of tumor (CT) and the invasive margin (IM) [14, 16]. A high density of immune marker positive lymphocytes in each region was recorded as a score. IS is a summation of the score of CD3+ and CD8+ TILs in the CT and IM, which is from 0 to 4. Then, all the patients could be divided into two groups—IS low group (0, 1, and 2) or high group (3, 4).

2.4. Statistics. All data were statistically analyzed by the Statistical Package for the Social Sciences, version 23.0 (SPSS Inc., Chicago, IL, USA). The correlation among clinicopathological features and mutation was calculated by a Chi-square test (for categorical variables) and Student t-test (for continuous variables). Overall survival was calculated by the Kaplan-Meier method. For identifying the independent prognostic factors for OS, the Cox proportional-hazards model was used for univariate and multivariate analyses. P value less than 0.05 was considered to be statistically significant.

3. Results

3.1. Basic Characteristics of the Recruited mCRC Patients. We analyzed the basic characteristics of the recruited mCRC patients (Table 1). We found RAS gene mutant tumors were more likely to develop in the right colon in comparison with RAS wild-type tumors (68.75% versus 31.09%, $P = 0.017$). PD-L1 was more likely to express in the rectum in comparison with colon (68.00 versus 25.71%, $P = 0.001$).

3.2. Survival Analysis Associated with RAS Status. We sequenced all coding exons of all three RAS isoforms in the 60 mCRCs at first. The Kaplan-Meier survival analysis demonstrated that there were no significant differences in OS between RAS ($P = 0.069$), KRAS ($P = 0.114$), mutation type and wild type (Figures 1(a) and 1(b)).

3.3. Prognostic Value of Immunoscore in mCRCs. The immunohistochemical results of the CD3 and CD8 were showed in Figure 2(a). IS is a summation of the score of CD3+ and CD8+ TILs in the CT and IM, which is from 0 to 4. Then, all the patients were divided into two groups—IS low group (0, 1, and 2) and high group (3, 4). The Kaplan-Meier analysis showed immunoscore (IS) was not significantly correlated with survival ($P = 0.799$) (Figure 2(b)).

Then, we divided these patients into two groups by IS. The Kaplan-Meier analysis shows RAS gene type was not significantly correlated with survival in each group ($P = 0.101$, $P = 0.387$, resp.). But, by univariate COX regression analysis, the P value and hazard ratios were 0.140 and 0.277 in IS-High group (Figures 2(c) and 2(d)).

3.4. Prognostic Value of PD-L1 Expression in mCRCs. The immunohistochemical results of the PD-L1 expression were showed in Figure 3(a). All the patients were divided into two groups with or without the expression of PD-L1. The Kaplan-Meier analysis showed the PD-L1 expression was not significantly correlated with survival ($P = 0.143$) (Figure 3(b)).

Then, we divided these 60 patients into another two groups by PD-L1 expression. The Kaplan-Meier analysis showed RAS gene type was not significantly correlated with survival in each group, either ($P = 0.287$, $P = 0.052$, resp.). But, by univariate COX regression analysis, the P value and hazard ratios were 0.080 and 0.24 in PD-L1-negative group (Figures 3(c) and 3(d)).

3.5. Univariate and Multivariable Analyses in mCRCs. We used the Cox proportional-hazards model to investigate the independent prognostic factors for OS in patients with mCRC (Table 2). The univariate analysis showed that the OS of patients with RAS mutation was worse than patients without RAS mutation (hazard ratio (HR): 0.473), though the P value is not significant ($P = 0.069$). In multivariate analysis, RAS mutation and PD-L1 expression in lymphocyte were demonstrated to be the independent negative prognostic factor for OS ($P = 0.044$, HR: 0.258, and 95% CI: 0.069–0.967;

Table 1: Basic characteristics of the recruited mCRC patients.

Characteristics	Total	*RAS mutation*			*Immunoscore*			*PD-L1 expression*		
		Mutation type	Wild type	*P* value	Low	High	*P* value	Negative	Positive	*P* value
Patients number (percentage)	60	26 (43.33%)	34 (56.67%)		38 (63.33%)	22 (36.67%)		34 (56.67%)	26 (43.33%)	
Age				0.341			0.325			0.207
Mean ± SD		59.64 ± 10.68	59.15 ± 10.51		59.84 ± 10.16	59.22 ± 11.53		59.41 ± 8.85	59.88 ± 12.69	
Sex				0.832			0.294			0.429
Male	43	19	24		29	14		23	20	
Female	17	7	10		9	8		11	6	
Location				0.054			0.851			0.002
Right	16	11	5	0.017	11	5	0.600	10	6	0.582
Left	19	7	12		12	7		16	3	
Rectum	25	8	17	0.134	15	10	0.651	8	17	0.001
Site of metastasis										
Liver	53	22	31	0.433	34	19	0.718	30	23	0.978
Lung	11	7	4	0.133	7	4	0.982	8	3	0.234
Others	4	2	2	0.781	2	2	0.567	2	2	0.781

Age was compared between two groups by using independent *t*-test; *P* values are calculated by using Fisher's exact test because less than 80% of the cells have an expected frequency of 5 or greater, or any cell has an expected frequency smaller than 1.0.

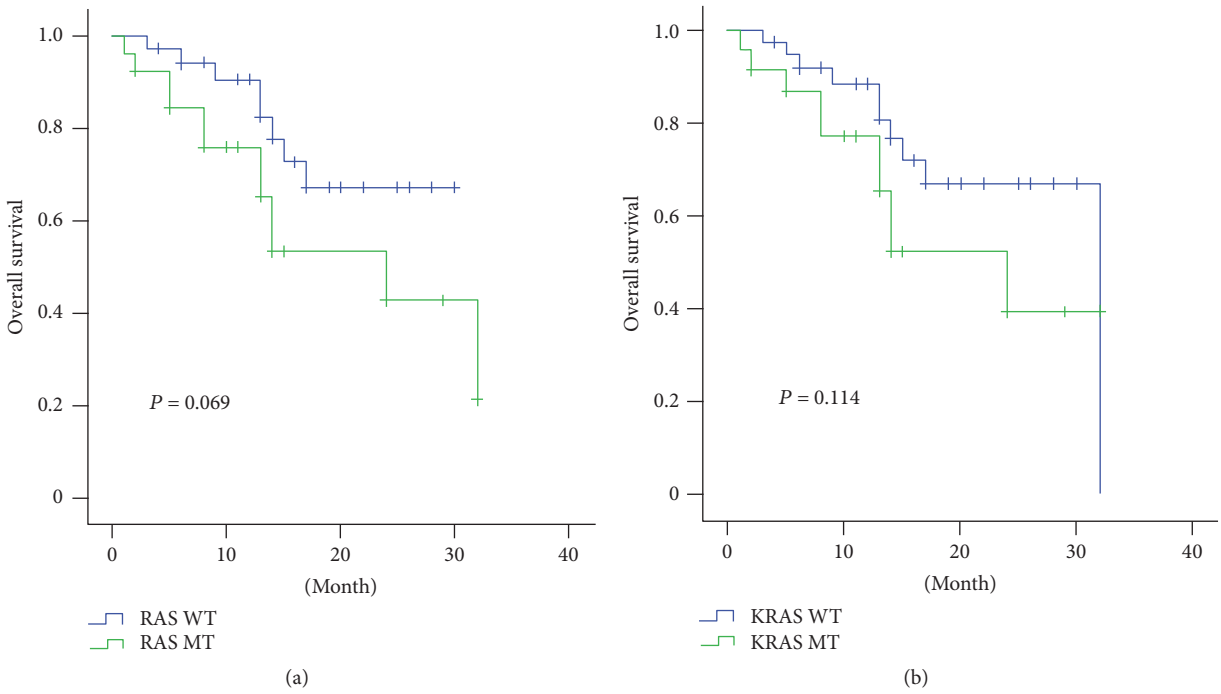


Figure 1: *Relationship of RAS status and overall survival in mCRC.* (a) Overall survival analysis to RAS statue of all the patients. (b) Overall survival analysis to KRAS statue of all the patients.

P = 0.048, HR: 0.276, and 95% CI: 0.077–0,988). And both IS and age had impressive influence on OS (HR: 2.681; HR: 2.127).

4. Discussion

In this study, we elucidated the prevalence of RAS mutations in Chinese mCRC patients, clarified the correlation between clinicopathological features and gene status, and investigated the prognostic value of tumor-infiltrating cells. So far, most clinical evidence about RAS and BRAF mutations in mCRC were originated from western countries. In this paper, we detected the frequency of RAS and KRAS mutation in 60 Chinese mCRC patients with palliative operation (53.33%, 38.33%). More recently, several reports have shown that exon 3 or 4 mutation of KRAS and exons 2–4 mutation of NRAS occurred in approximately 10 percent of mCRC patients with KRAS exon 2 wild-type tumors. Our data showed that the frequency of patients with KRAS exon 2 mutant tumors is similar.

As previously reported, the presence of BRAF mutations in CRC was always a strongly poor prognostic marker for clinical outcome. And patients with BRAF mutant are often refractory to systematic chemotherapy [17]. However, there was no identical conclusion about the correlation between the RAS mutation and the prognosis in mCRC patients. Previously, research showed that there was insufficient evidence to definitively state that patients with RAS mutations mCRC could benefit from bevacizumab combined with chemotherapy as first-line treatment [18]. Recently, several studies have demonstrated that immunoscore (IS) has high prognostic utility, which could be demonstrated as the density of CD3+ and CD8+ lymphocytes in the tumor center (CT) and invasive margin (IM) [16, 19, 20]. Moreover, it has been reported that the IS method is much better while compared to the current tumor-node-metastases (TNM) staging system, especially in colon cancers [21]. In a recent report, Lea et al. described the limitations of the current TNM staging system in predicting the outcome of patients with CRC [22]. They suggested that the immune cell density in the stromal environment could be a better prognostic marker. This suggestion was also confirmed by Mlecnik et al. [23]. Furthermore, the multivariate survival analysis conducted by Anitei et al. confirmed that the IS system has stronger prognostic value than the TNM staging system [24]. In this study, all the patients were mCRC with palliative operation and we demonstrated the prognostic value of the IS method. We divided all the patients to low IS (0, 1, and 2) and high group (3, 4). Our study demonstrated that patients without RAS mutation have a better prognostic in the higher density of CD3+ and CD8+ lymphocytes group. Most of the studies have demonstrated that dense infiltration of CD3+ and CD8+ lymphocytes is associated with less aggressive clinic-pathological features and a better prognosis [24, 25]. Hence, the IS system could be a robust prognostic factor that is assessable for mCRC patients without RAS mutation.

Previous study suggested that the activation of the PD-1/PD-L1 signaling pathway created an immunosuppressive tumor microenvironment for tumors to escape from immune clearance [26]. Thus, blockade of the PD-1/PD-L1 function provided a potential strategy for cancer immunotherapy. Many clinical trials have been conducted to show the clinical benefit of various types of tumors from anti-PD-1/PD-L1 immunotherapy, such as malignant melanoma,

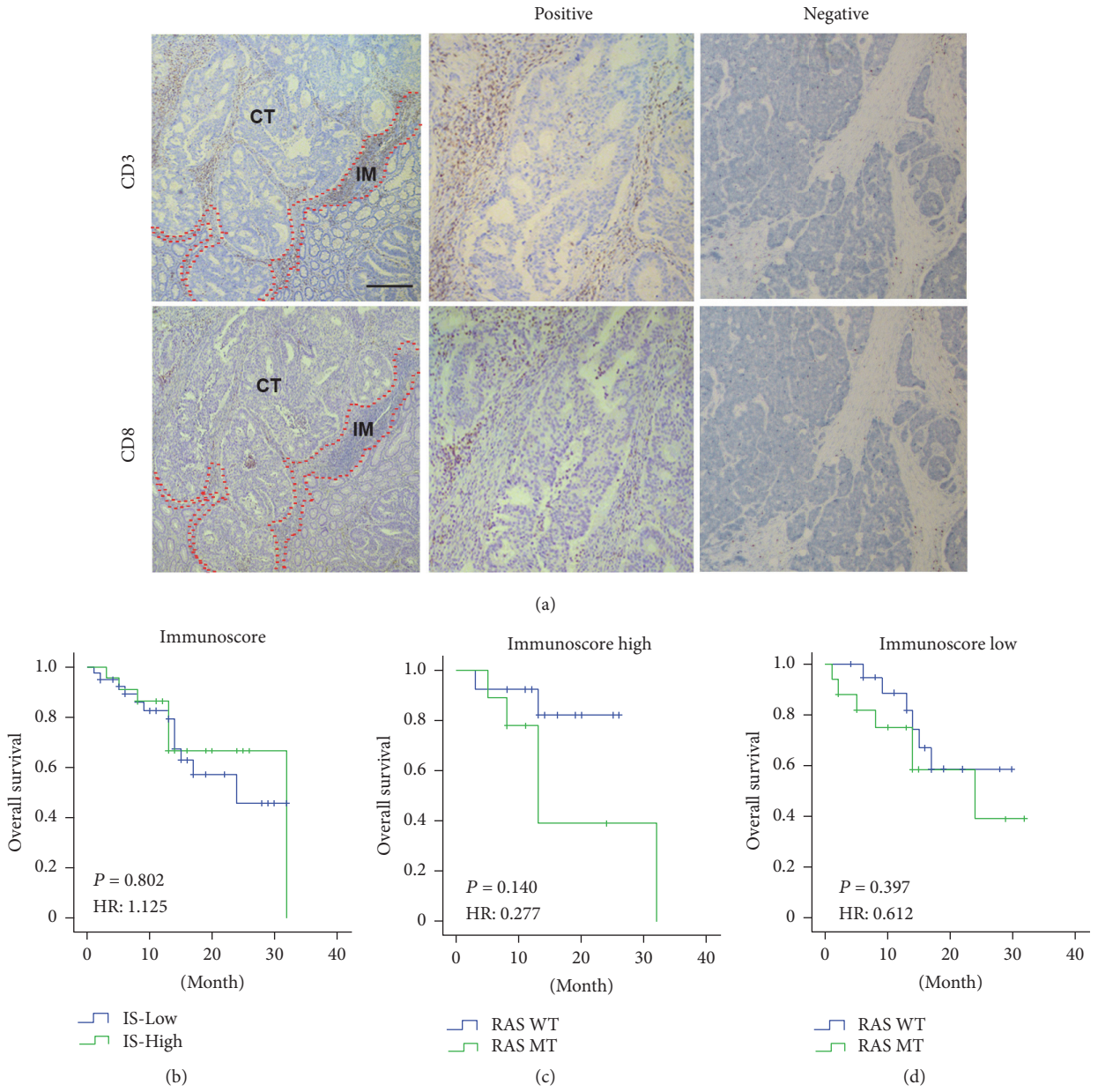


Figure 2: *Prognostic value of immunoscore in mCRCs.* (a) The immunohistochemical results of the CD3 and CD8 in the CT and IM of the primary tumor. (b) The Kaplan-Meier survival curve according to IS. (c) The Kaplan-Meier survival curve according to RAS statue in IS-High patients. (d) The Kaplan-Meier survival curve according to RAS statue in IS low patients.

non-small cell lung cancer, and renal cell carcinoma [27, 28]. A recent phase II trial reported that mismatch-repair status could predict a survival benefit during blockade of the immune checkpoint system in CRC patients [29]. Interestingly, several studies found that PD-L1 expression was also correlated to MSI status [30]. In our study, we found that high IS correlated with prolonged OS and was a good independent prognostic indicator in RAS wild-type mCRC patients. According to other research, high PD-1 expression has been correlated with improved response to immune checkpoint inhibitors, compared with low PD-L1 expression. Furthermore, PD-L1 expression on the peritumor cells may be correlated with improved response to immune checkpoint inhibitors. In addition, the high mutational frequency found within tumors raises the possibility that T cells may preferentially invade tumors in patients whose T cells recognize mutated epitopes found within the tumor tissue [31]. These findings suggest that PD-L1 expression is a useful and reproducible tool for predicting survival for mCRC patients. In our study, we divided the 60 mCRC patients into two groups according to the percent of PD-L1 expression in tumor cell and lymphocytes. The Kaplan-Meier analysis showed that there was a better prognostic with PD-L1 expression in wild-type RAS patients. We found that the

Table 2: Univariate and multivariate analyses of OS in 60 mCRC patients.

Variables	Univariate analysis			Multivariate analysis		
	P value	HR	95% CI	*P* value	HR	95% CI
Age (≥60)	0.079	2.255	0.909–5.597	0.166	2.127	0.731–6.188
Location (left/right)	0.714	0.826	0.298–2.292	0.534	0.631	0.148–2.691
RAS mutation	0.109	0.473	0.189–1.181	0.044	0.258	0.069–0.967
Histology	0.228	0.551	0.209–1.453	0.467	0.643	0.195–2.114
Nerve invasion	0.587	1.293	0.512–3.265	0.954	0.969	0.334–2.812
Vascular invasion	0.719	1.203	0.440–3.285	0.613	0.734	0.221–2.433
Immunoscore	0.802	1.125	0.447–2.831	0.127	2.681	0.756–9.507
PD-L1	0.160	0.531	0.219–1.284	0.048	0.276	0.077–0.988

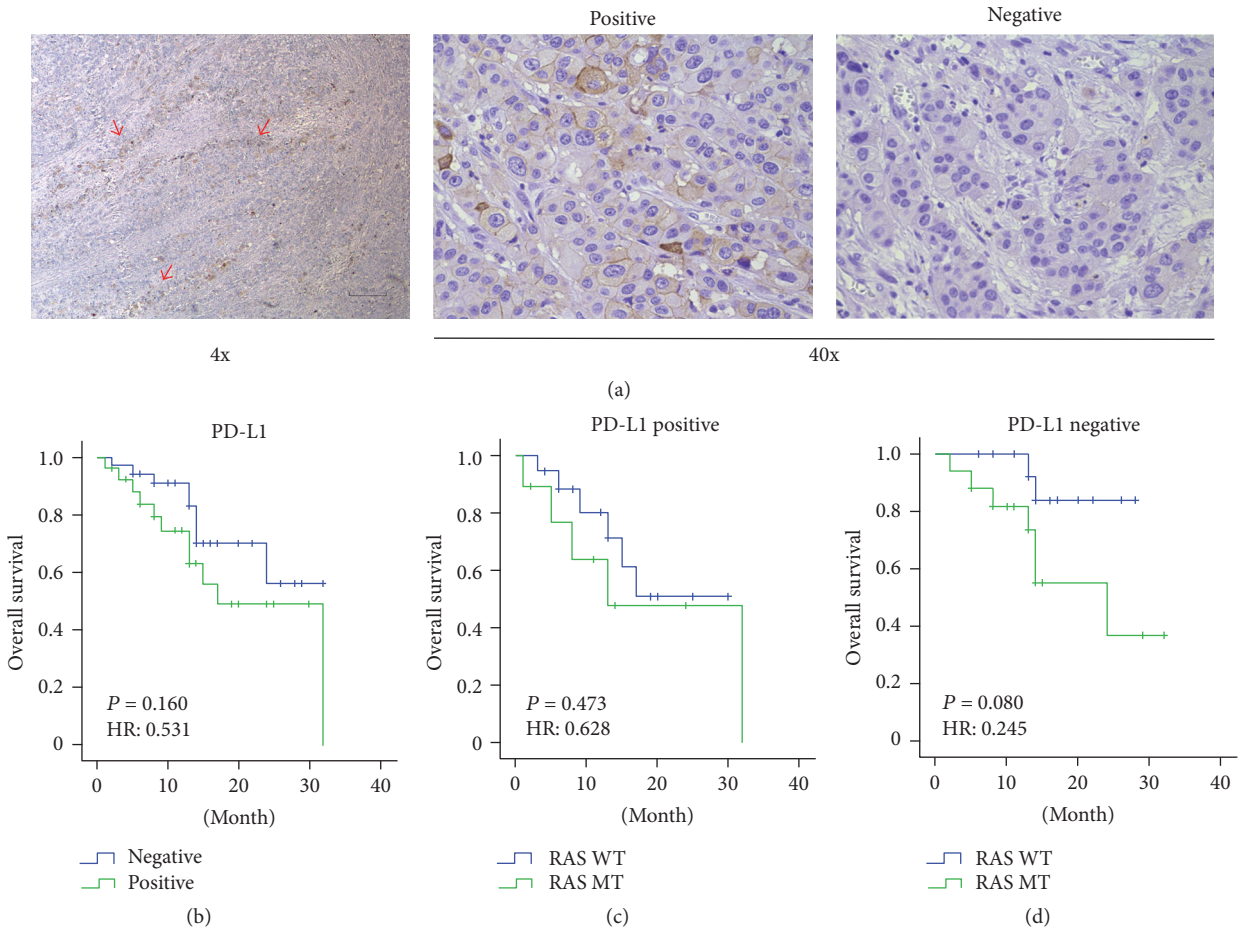


Figure 3: *Prognostic value of PD-L1 expression in mCRCs.* (a) The PD-L1 expression of the primary tumor. (b) The Kaplan-Meier survival curve according to PD-L1 expression. (c) The Kaplan-Meier survival curve according to RAS statue in patients with PD-L1 expression. (d) The Kaplan-Meier survival curve according to RAS statue in patients without PD-L1 expression.

PD-L1 expression was the independent negative prognostic factor for OS in multivariate analysis ($P = 0.048$, HR: 0.276, and 95% CI: 0.077–0.988).

5. Conclusions

In conclusion, for the mCRC patients with palliative operation and negative PD-L1 expression, the RAS mutation is a negative prognostic factor. And the RAS mutation maybe a potential negative prognostic factor for the mCRC patients with palliative operation and high immunoscore. All the results suggested that, combined with RAS status, IS and PD-L1 expression may be the prognostic indicators for mCRC patients with palliative operation. This will provide a better prognostic marker for the treatment of mCRC patients without radical operation.

Abbreviations

CRC: Colorectal cancer
CT: Core of tumor
HR: Hazard ratio
IM: Invasive margin
IS: Immunoscore
mCRC: Metastatic colorectal cancer
MT: Mutant type
OS: Overall survival
PD-L1: Programmed cell death-ligand 1
TILs: Tumor-infiltrating lymphocytes
TNM: Tumor node metastases
WT: Wild type.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Authors' Contributions

Ruiqi Liu, Ke Peng, and Yiyi Yu have contributed equally to this work.

Acknowledgments

This study was funded by National Natural Science Foundation of China (Grants nos. 81772511, 81602038, and 81502003).

References

[1] F. Tao, J. Lv, W. Wang, and K. Jin, "Current management of colorectal hepatic metastasis," *International Journal of Clinical and Experimental Medicine*, vol. 15, no. 8, pp. 19850–19858, 2015.

[2] W. Chen, R. Zheng, P. D. Baade et al., "Cancer statistics in China, 2015," *CA: A Cancer Journal for Clinicians*, vol. 66, no. 2, pp. 115–132, 2016.

[3] M. Zabala, P. Alzuguren, C. Benavides et al., "Evaluation of bioluminescent imaging for noninvasive monitoring of colorectal cancer progression in the liver and its response to immunogene therapy," *Molecular Cancer*, vol. 8, article no. 2, 2009.

[4] S. Negru, E. Papadopoulou, A. Apessos et al., "KRAS, NRAS and BRAF mutations in Greek and Romanian patients with colorectal cancer: A cohort study," *BMJ Open*, vol. 4, no. 5, Article ID e004652, 2014.

[5] E. Van Cutsem, M. Peeters, S. Siena et al., "Open-label phase III trial of panitumumab plus best supportive care compared with best supportive care alone in patients with chemotherapy-refractory metastatic colorectal cancer," *Journal of Clinical Oncology*, vol. 25, no. 13, pp. 1658–1664, 2007.

[6] J. A. McCubrey, L. S. Steelman, S. L. Abrams et al., "Roles of the RAF/MEK/ERK and PI3K/PTEN/AKT pathways in malignant transformation and drug resistance," *Advances in Enzyme Regulation*, vol. 46, no. 1, pp. 249–279, 2006.

[7] M. Scaltriti and J. Baselga, "The epidermal growth factor receptor pathway: A model for targeted therapy," *Clinical Cancer Research*, vol. 12, no. 18, pp. 5268–5272, 2006.

[8] G. Di Caro, F. Marchesi, L. Laghi, and F. Grizzi, "Immune cells: plastic players along colorectal cancer progression," *Journal of Cellular and Molecular Medicine*, vol. 17, no. 9, pp. 1088–1095, 2013.

[9] N. A. Giraldo, E. Becht, R. Remark, D. Damotte, C. Sautès-Fridman, and W. H. Fridman, "The immune contexture of primary and metastatic human tumours," *Current Opinion in Immunology*, vol. 27, no. 1, pp. 8–15, 2014.

[10] W. H. Fridman, J. Galon, M.-C. Dieu-Nosjean et al., "Immune infiltration in human cancer: prognostic significance and disease control," *Current Topics in Microbiology and Immunology*, vol. 344, pp. 1–24, 2011.

[11] J. Galon, F. PagΦs, and F. M. Marincola, "Cancer classi cation using the Immunoscore: a worldwide task force," *Journal of Translational Medicine*, vol. 10, no. 205, 2012.

[12] A. Gabrielson, Y. Wu, H. Wang et al., "Intratumoral CD3 and CD8 T-cell densities associated with relapse-free survival in HCC," *Cancer Immunology Research*, vol. 4, no. 5, pp. 419–430, 2016.

[13] B. Mlecnik, G. Bindea, H. K. Angell et al., "Integrative Analyses of Colorectal Cancer Show Immunoscore Is a Stronger Predictor of Patient Survival Than Microsatellite Instability," *Immunity*, vol. 44, no. 3, pp. 698–711, 2016.

[14] C. Böger, H.-M. Behrens, M. Mathiak, S. Krüger, H. Kalthoff, and C. Röcken, "PD-L1 is an independent prognostic predictor in gastric cancer of Western patients," *Oncotarget* , vol. 7, no. 17, pp. 24269–24283, 2016.

[15] Y. Kwak, J. Koh, D. Woo, and et al., "KimImmunoscore encompassing CD3+ and CD8+ T cell densities in distant metastasis is a robust prognostic marker for advanced colorectal cancer," *Oncotarget*, vol. 7, no. 49, pp. 81778–81790, 2016.

[16] Y. Kwak, J. Koh, D.-W. Kim, S.-B. Kang, W. H. Kim, and H. S. Lee, "Immunoscore encompassing CD3+ and CD8+ T cell densities in distant metastasis is a robust prognostic marker for advanced colorectal cancer," *Oncotarget*, vol. 6, no. 7, pp. 81778–81790, 2016.

[17] J. Galon, B. Mlecnik, G. Bindea et al., "Towards the introduction of the 'Immunoscore' in the classi cation of malignant tumours," *The Journal of Pathology*, vol. 232, no. 2, pp. 199–209, 2014.

[18] S. Siena, F. Rivera, J. Taieb et al., "Survival Outcomes in Patients With RAS Wild Type Metastatic Colorectal Cancer Classified According to Köhne Prognostic Category and BRAF Mutation Status," *Clinical Colorectal Cancer*, 2017.

[19] M. Zhou, P. Yu, J. Qu et al., "Efficacy of Bevacizumab in the First-Line Treatment of Patients with RAS Mutations Metastatic Colorectal Cancer: A Systematic Review and Network Meta-Analysis," *Cellular Physiology and Biochemistry*, vol. 40, no. 1-2, pp. 361–369, 2016.

[20] W. H. Fridman, F. Pagès, C. Sautès-Fridman, and J. Galon, "The immune contexture in human tumours: impact on clinical outcome," *Nature Reviews Cancer*, vol. 12, no. 4, pp. 298–306, 2012.

[21] J. Galon, F. Pages, and F. M. Marincola, "Cancer classi cation using the Immunoscore: a worldwide task force," *Journal of Translational Medicine*, vol. 3, no. 10, 2012.

[22] D. Lea, S. Haland, H. R. Hagland, and K. Soreide, "Accuracy of TNM staging in colorectal cancer: A review of current culprits, the modern role of morphology and stepping-stones for improvements in the molecular era," *Scandinavian Journal of Gastroenterology*, vol. 49, no. 10, pp. 1153–1163, 2014.

[23] B. Mlecnik, M. Tosolini, A. Kirilovsky et al., "Histopathologic-based prognostic factors of colorectal cancers are associated

with the state of the local immune reaction," *Journal of Clinical Oncology*, vol. 29, no. 6, pp. 610–618, 2011.

[24] M.-G. Anitei, G. Zeitoun, B. Mlecnik et al., "Prognostic and predictive values of the immunoscore in patients with rectal cancer," *Clinical Cancer Research*, vol. 20, no. 7, pp. 1891–1899, 2014.

[25] J. Galon, A. Costes, F. Sanchez-Cabo et al., "Type, density, and location of immune cells within human colorectal tumors predict clinical outcome," *Science*, vol. 313, no. 5795, pp. 1960–1964, 2006.

[26] E. Sato, S. H. Olson, and J. Ahn, "Intraepithelial $CD8^+$ tumor-infiltrating lymphocytes and a high $CD8^+$/regulatory T cell ratio are associated with favorable prognosis in ovarian cancer," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 102, no. 51, pp. 18538–18543, 2005.

[27] S. Spranger, R. M. Spaapen, Y. Zha et al., "Up-regulation of PD-L1, IDO, and T(regs) in the melanoma tumor microenvironment is driven by CD8(+) T cells," *Science Translational Medicine*, vol. 5, no. 200, p. 200ra116, 2013.

[28] E. de Guillebon, P. Roussille, E. Frouin, and D. Tougeron, "Anti program death-1/anti program death-ligand 1 in digestive cancers," *World Journal of Gastrointestinal Oncology*, vol. 7, no. 8, pp. 95–101, 2015.

[29] K. Muro, H. C. Chung, V. Shankaran et al., "Pembrolizumab for patients with PD-L1-positive advanced gastric cancer (KEYNOTE-012): a multicentre, open-label, phase 1b trial," *The Lancet Oncology*, vol. 17, no. 6, pp. 717–726, 2016.

[30] L. A. Diaz, D. T. Le, and et al., "PD-1 blockade in tumors with mismatch-repair deficiency," *The New England Journal of Medicine*, vol. 12, no. 373, p. 1979, 2015.

[31] J. M. Taube, "Unleashing the immune system: PD-1 and PD-Ls in the pre-treatment tumor microenvironment and correlation with response to PD-1/PD-L1 blockade," *OncoImmunology*, vol. 3, no. 11, pp. e963413–e963413-3, 2014.

Hindawi
BioMed Research International
Volume 2018, Article ID 2914280, 11 pages
https://doi.org/10.1155/2018/2914280

Research Article

A Risk Stratification Model for Lung Cancer Based on Gene Coexpression Network and Deep Learning

Hongyoon Choi [1] **and Kwon Joong Na** [2,3]

[1] *Cheonan Public Health Center, Chungnam, Republic of Korea*
[2] *Department of Community Health, Korea Health Promotion Institute, Seoul, Republic of Korea*
[3] *Department of Clinical Medical Sciences, Seoul National University, College of Medicine, Seoul, Republic of Korea*

Correspondence should be addressed to Hongyoon Choi; chy1000@gmail.com and Kwon Joong Na; kjna85@gmail.com

Received 13 October 2017; Revised 7 December 2017; Accepted 11 December 2017; Published 16 January 2018

Academic Editor: Jialiang Yang

Risk stratification model for lung cancer with gene expression profile is of great interest. Instead of previous models based on individual prognostic genes, we aimed to develop a novel system-level risk stratification model for lung adenocarcinoma based on gene coexpression network. Using multiple microarray, gene coexpression network analysis was performed to identify survival-related networks. A deep learning based risk stratification model was constructed with representative genes of these networks. The model was validated in two test sets. Survival analysis was performed using the output of the model to evaluate whether it could predict patients' survival independent of clinicopathological variables. Five networks were significantly associated with patients' survival. Considering prognostic significance and representativeness, genes of the two survival-related networks were selected for input of the model. The output of the model was significantly associated with patients' survival in two test sets and training set ($p < 0.00001$, $p < 0.0001$ and $p = 0.02$ for training and test sets 1 and 2, resp.). In multivariate analyses, the model was associated with patients' prognosis independent of other clinicopathological features. Our study presents a new perspective on incorporating gene coexpression networks into the gene expression signature and clinical application of deep learning in genomic data science for prognosis prediction.

1. Introduction

Risk stratification based on gene expression profiles is of major biomedical interest in lung cancer research [1–6]. Previous studies developed risk stratification models that mostly focused on individual prognostic genes. However, these studies have not fully considered the nature of biological networks and their systematic properties. Since it is more evident that biological processes are derived from numerous interactions between many cellular components, gene network analysis could provide valuable information of cancer pathogenesis [7]. Among the various biological networks, gene coexpression network has some strengths: not relying on prior information about genes, avoiding biologically wrong assumptions about independence of gene expression levels, and alleviating multiple testing problems [8].

Lung cancer, mainly, non-small-cell lung cancer, is one of the most common cancers and is the leading cause of cancer-related death worldwide [9, 10]. Currently, TNM staging system is a universal guideline for prognosis prediction and treatment decision. However, heterogeneous molecular features of lung cancer require diverse adjuvant treatment options and lead to different prognosis even in the same stage [11]. Hence, there has been a constant need for developing better risk stratification models to predict accurate prognosis and to improve cancer-related survival.

The main objectives of this study were (1) to identify survival-related gene coexpression network modules (2) and to propose a deep learning- (DL-) based risk stratification model reflecting survival-related network modules. Using public microarray datasets from the Gene Expression Omnibus (GEO), we identified survival-related network modules

of lung adenocarcinoma. Subsequently, we constructed DL-based prognostic score using representative genes of survival-related network modules and it showed great prognostic property in all cohorts.

2. Materials and Methods

2.1. Gene Expression Data. Total eleven microarray datasets from NCBI GEO were included in the study. Two datasets with survival information were set as independent test sets and the others were merged and set as training set (Supplementary Table 1). Detailed preprocessing methods are described in Supplementary Methods (available here).

2.2. Weighted Gene Coexpression Network Construction from the Training Set. We used weighted gene coexpression network analysis (WGCNA) package [8, 12] to build a weighted gene coexpression network from the training set. We created a correlation matrix on the basis of Pearson's correlation coefficient for all pairwise genes across all samples. The power, the key parameter for weighted network, was selected to optimize both the scale-free topology and sufficient node connectivity and we chose a threshold of 6 in this study (Supplementary Figure 1). The correlation matrix was transformed into adjacency matrix using the power function, and pairwise topological overlap (TO) between genes was calculated. We identified network modules using hierarchical clustering method with TO dissimilarity as the distance measure. The modules were detected using dynamic tree cut algorithm [13], defining height cutoff value of 0.99, deep split as 2, and minimum module size cutoff value of 30. Genes that were not assigned to any module were classified to color gray (Figure 1).

2.3. Identification and Validation of Survival-Related Network Modules. For each module, we summarized the module expression profile by module eigengene (ME), which is the first principal component of the expression matrix of the corresponding module. We used ME as the representative of each module to evaluate association with overall survival (OS). The survival-related network modules were identified using Cox regression analysis in the training set. For validation, the same genes included in the network construction were extracted from each test set. ME was calculated based on the expression profile of each test set, and the association between ME and OS was evaluated using Cox regression analysis to see whether the modules identified from the training set are also associated with OS in each test set. The modules with uncorrected p value under 0.05 were regarded as significant survival-related network modules. We functionally annotated all survival-related network modules with gene ontology biological process terms using hypergeometric test (Supplementary Methods).

2.4. DL-Based Risk Stratification Model. To simplify risk stratification model, we selected representative genes of the survival-related modules for model construction. Representativeness of a gene was measured by gene module membership (GMM), a correlation coefficient between gene expression profile and ME of given module. Expression profiles of representative genes were used for the input of the DL because they were expected to preserve coexpression patterns and to reflect the systematic properties of survival-related network modules. Convolutional neural network (CNN) was specifically used to extract gene expression patterns of modules. It finally produced gene network prognostic score (NetScore). Details of selection of representative genes and architecture of DL framework are described in Supplementary Methods.

The DL-based risk stratification model was generated using patients' data of the training set. Parameters related to training of the neural network including number of layers, nodes, training epoch and learning rate were determined by 5-fold cross-validation. Training set was randomly divided into 5 subsets. At each step, a single subset was left for testing and other four subsets were used for training. The performance of the model was measured by Harrell's C-index of the final output score of the model [14]. The optimal parameters were selected according to the maximum average C-index across the 5-fold of the loop. The predictive value of NetScore was independently validated in two test sets. C-index for each test set was also evaluated.

2.5. Survival Analysis Using NetScore in All Cohorts. Prognostic property of NetScore as a continuous variable was evaluated by univariate Cox analysis. To define risk groups, NetScore was dichotomized using the median value in each cohort. Kaplan-Meier method was used to assess survival rates according to the risk groups and survival rate differences were assessed with the log-rank test. Additionally, independent prognostic value of NetScore was assessed by multivariate and subgroup analysis. Multivariate Cox analysis was performed using clinical and pathological variables as well as NetScore. Subgroups were divided on the basis of clinical and pathological features, and univariate Cox analysis of NetScore was performed in each subgroup.

3. Results

3.1. Gene Coexpression Network Modules from the Training Set. We aimed at developing a risk stratification model based on gene coexpression networks (Figure 1(a)). The networks were constructed from the training set which consists of microarray data of 510 lung adenocarcinoma samples. The clinicopathological features of all samples from the training set are detailed in Supplementary Table 2. Using WGCNA, 23 coexpression network modules were identified from the training set (Figure 1(b)). The relationship between modules is visualized with hierarchical clustering dendrogram and heatmap of the corresponding ME (Supplementary Figure 2).

3.2. Identification of Survival-Related Modules from the Training Set and Validation in Test Sets. Total five modules were significantly associated with OS (Figure 1(c)): red ($p < 0.0001$), turquoise ($p = 0.018$), magenta ($p = 0.029$), black ($p = 0.043$), and light green ($p = 0.044$). To validate the survival-related modules, we conducted survival analysis in two independent test sets (GSE31210 as test set 1 and GSE30219 as test set 2; $n = 226$ and 84, resp.). Consequently,

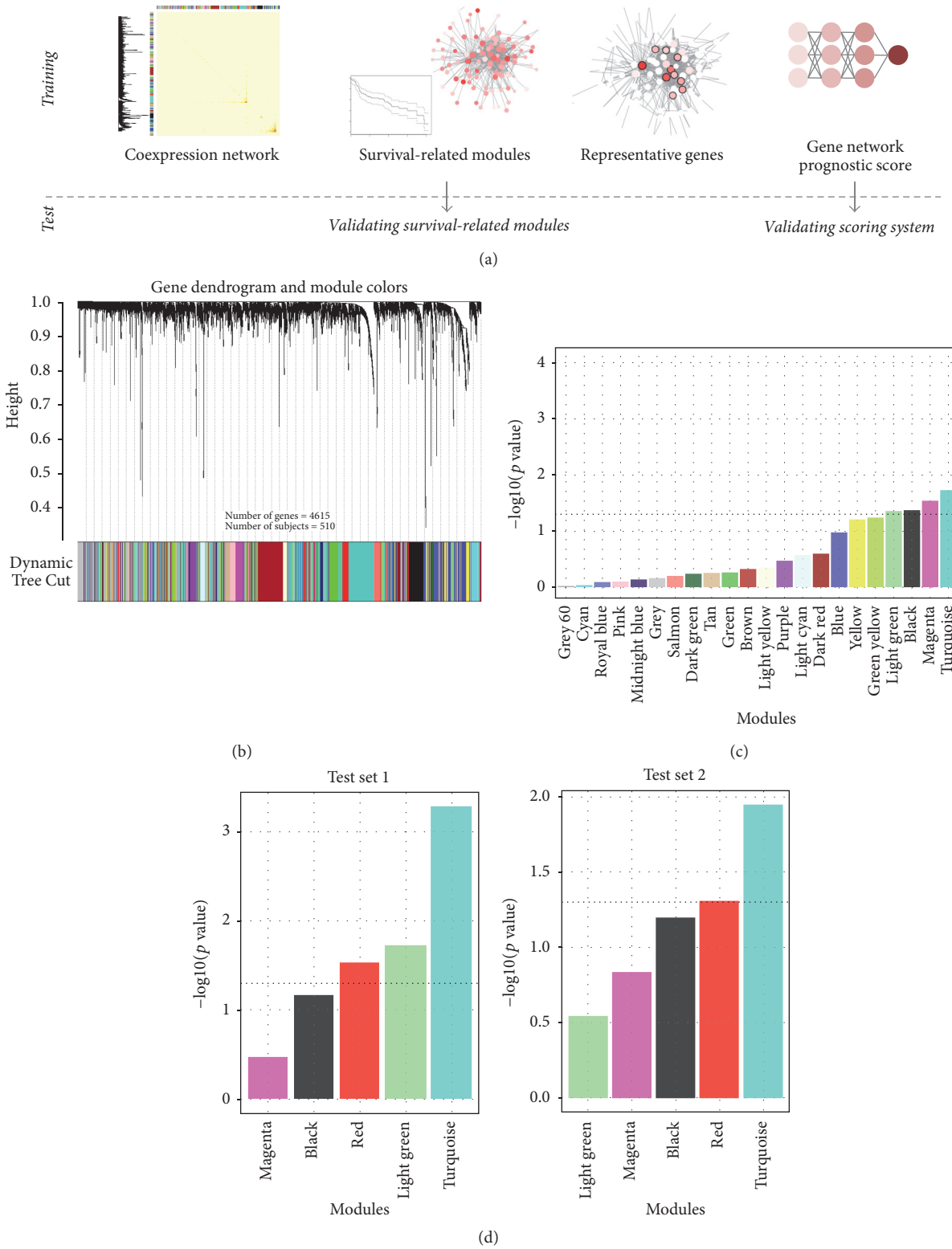


FIGURE 1: *Gene coexpression network construction and survival-related modules identification*. (a) A schematic diagram summarizing our risk stratification modeling strategy. Gene coexpression network was constructed from the training set. Gene network modules were extracted based on topological overlap. Survival-related modules were identified from the training set and validated in the two test sets. We selected representative genes from survival-related modules, and built network-based prognostic scoring system using deep learning. (b) Gene dendrogram and modules identified by weighted gene coexpression network analysis from the training set. Modules were labeled with different colors. (c) Univariate Cox regression analysis of module eigengene in the training set was performed. Module eigengene is a representative expression value of genes of each module calculated by the principal component analysis. The dotted line represents cutoff value (p value = 0.05) for significance, and five modules were identified as survival-related network modules. (d) Survival-related network modules were validated in the two test sets using Cox regression analysis. Three modules from test set 1 and two modules from test set 2 were significantly associated with overall survival.

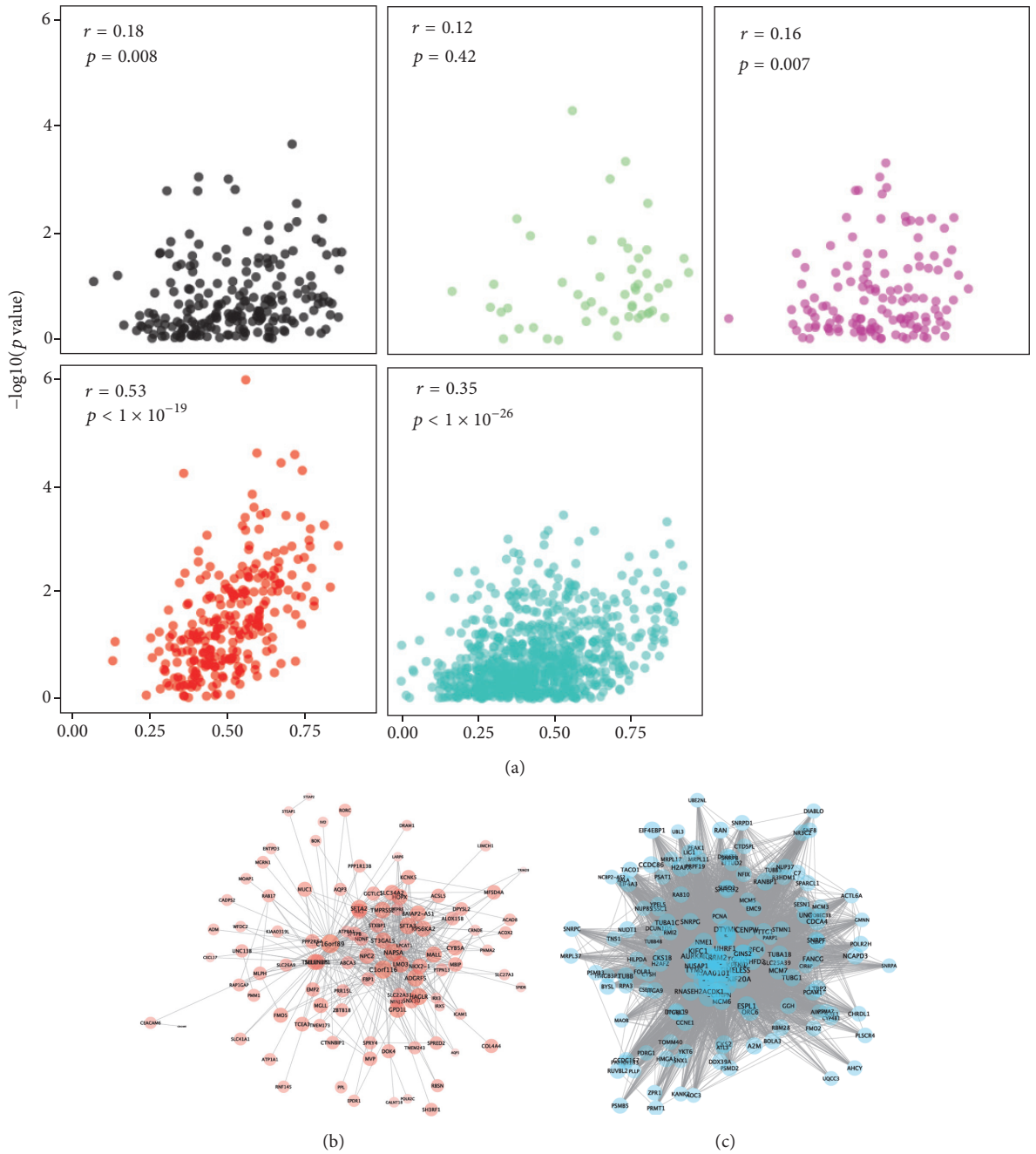
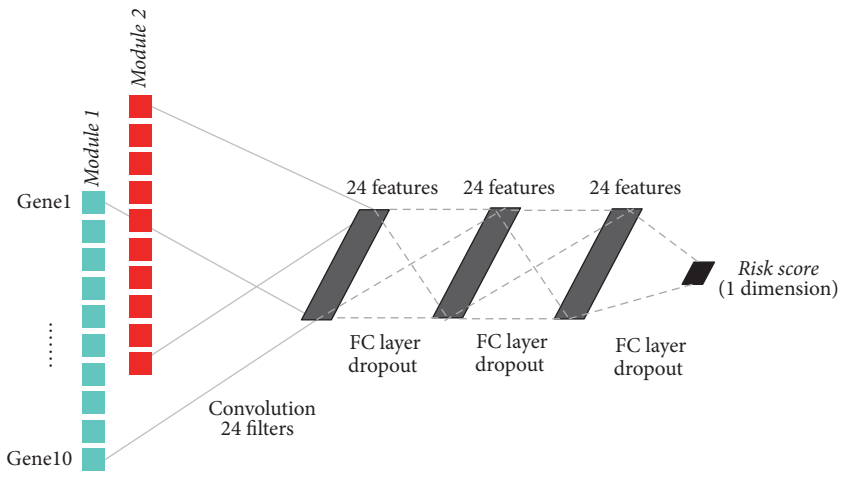


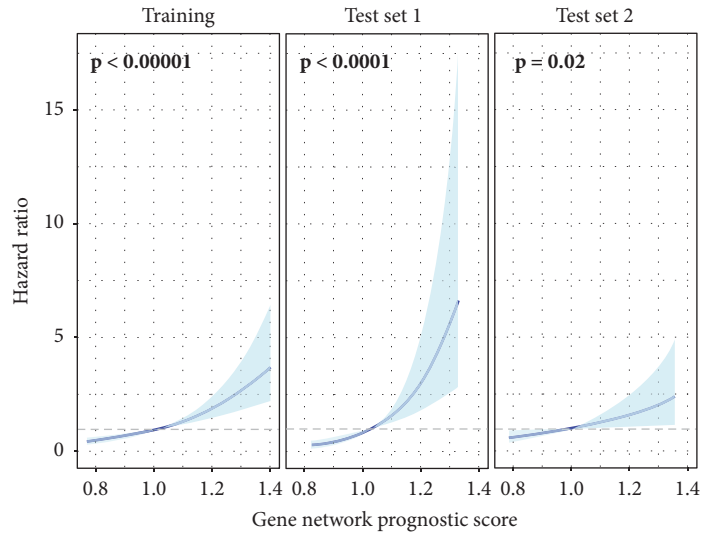
Figure 2: *Selection of representative genes of survival-related network modules*. (a) To construct risk stratification model, representative genes were selected according to the gene module membership. Gene module membership was correlated with the significance of association between individual gene expression and survival. y-axis represents statistical significance calculated by univariate Cox analysis of individual genes. A strong correlation was found in the red and turquoise modules ($r = 0.53$ and $p < 1 \times 10^{-19}$ for red module; $r = 0.35$ and $p < 1 \times 10^{-23}$ for turquoise module). Coexpression networks of red (b) and turquoise (c) modules were visualized. Note that 160 genes among 880 genes of turquoise module and their connections were shown. 160 genes were selected according to the gene module membership. Size of nodes is proportional to gene module membership.

turquoise ($p = 0.0005$), light green ($p = 0.019$), and red ($p = 0.030$) modules in test set 1 and turquoise ($p = 0.011$) and red ($p = 0.049$) modules in test set 2 were significantly associated with OS (Figure 1(d)).

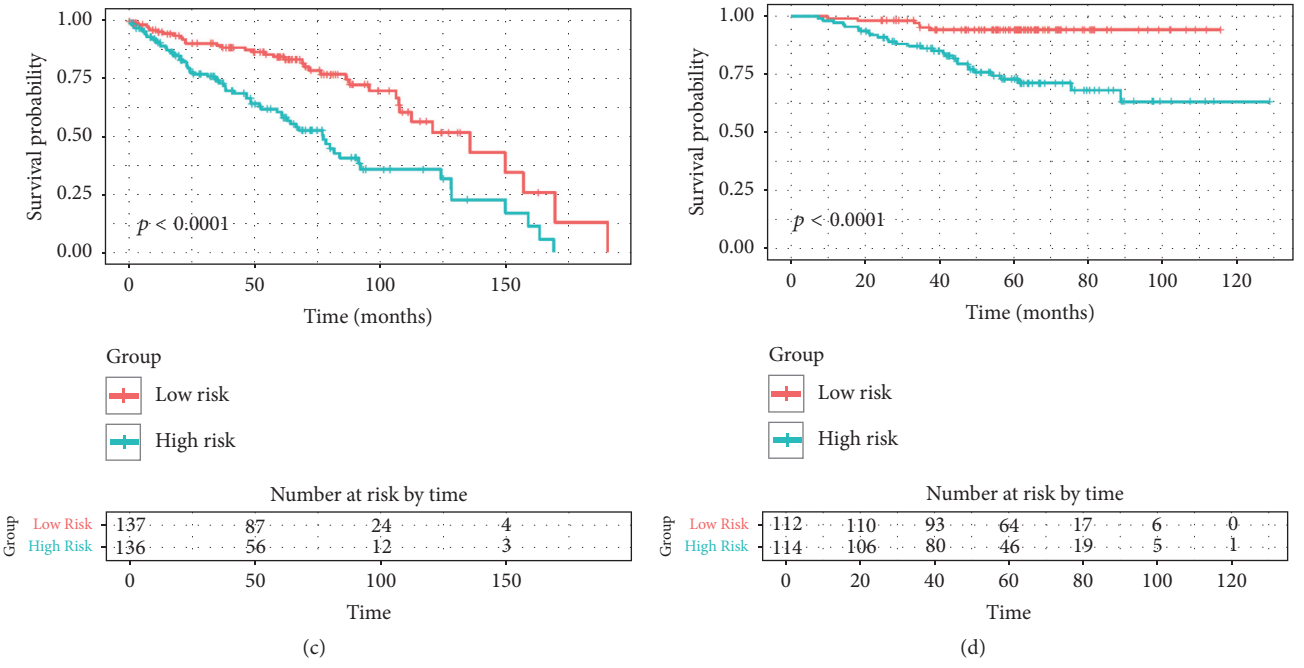
The networks of two common survival-related network modules (red and turquoise) are presented in Figures 2(a) and 2(b). The significantly enriched gene ontology terms of the red module included "organic acid catabolic process,"



(a)

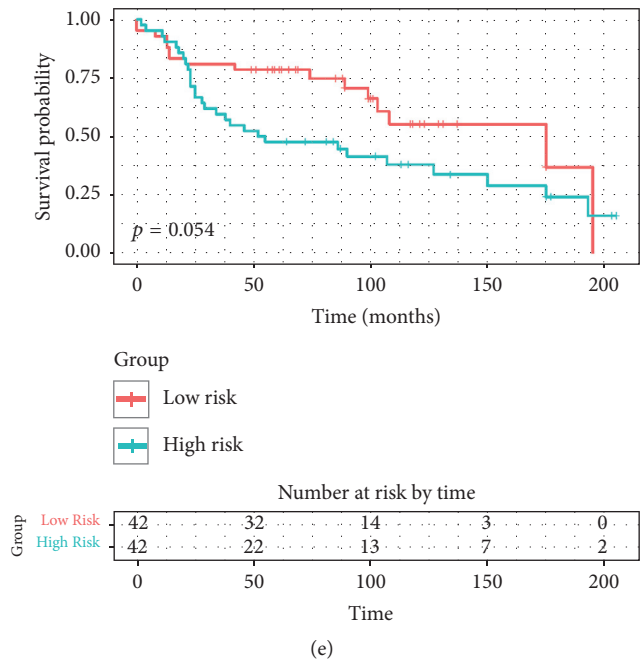


(b)



(c) (d)

Figure 3: Continued.



(e)

Figure 3: *Risk stratification model using representative genes of survival-related network modules*. (a) To construct risk stratification model, deep convolutional neural network was used. Input data were expression value of top 10 genes from each of red and turquoise module. The first layer consists of one-dimensional convolutional filters which extract gene expression patterns of each module. Three additional fully connected (FC) layers were followed and connected to the output score gene network prognostic score (NetScore). Detailed training process and architecture of the neural network are described in Supplementary Methods. (b) Univariate Cox regression analysis of NetScore as a continuous variable was performed in the training and two test sets. It shows significant association between the score and overall survival in all sets. The blue line represents hazard ratio for overall survival and the blue area represents 95% confidence interval. (c–e) Overall survival of dichotomized group according to NetScore was depicted by Kaplan-Meier survival curve. The statistical difference was tested by log-rank test. The high-risk group showed worse survival in the training set (c) and test set 1 (d) with statistical significance. The high-risk group of the test set 2 (e) also showed worse prognosis though the difference did not reach statistical significance.

"carboxylic acid catabolic process," and "small molecule catabolic process," and the turquoise module included "DNA strand elongation involved in DNA replication," "mitotic cell cycle phase transition," "DNA-dependent DNA replication" (Supplementary Table 3).

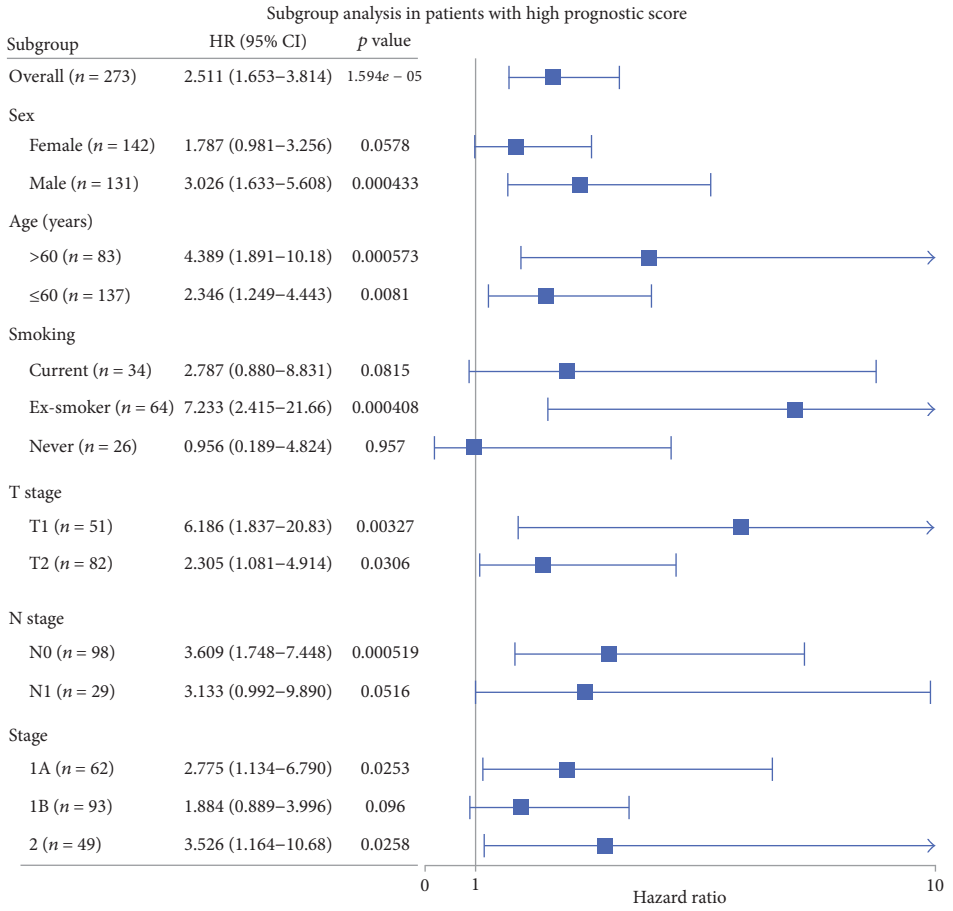
3.3. DL-Based Risk Stratification Model Using Representative Genes of Survival-Related Module. By measuring the correlation between gene significance for OS (p value) and GMM in each survival-related module, we identified two modules demonstrating high correlation with statistical significance ($r = 0.53$, $p < 1 \times 10^{-19}$ and $r = 0.35$, $p < 1 \times 10^{-26}$ for red and turquoise module, resp.; Figure 2(c)). Based on the strong correlation, we could assume that the genes with high representativeness measured by GMM have high significance for OS and are the most important elements of the module; therefore, we selected top 10 genes according to GMM from the red and turquoise modules for the DL-based risk stratification model construction (Supplementary Figure 3).

The expression profiles of selected 20 genes were used as input data of the risk stratification model (Figure 3(a)). NetScore, the final output of our model, was significantly associated with OS in the training and two test sets (Figure 3(b)) ($p < 0.00001$, $p < 0.0001$ and $p = 0.02$ for training set and test sets 1 and 2, resp.). Subjects were divided into two groups, high- and low-risk groups, according to the median value of NetScore in each cohort. The high-risk group was significantly associated with OS in the training set ($p < 0.0001$; Figure 3(c)) and in test set 1 ($p < 0.0001$; Figure 3(d)). A trend of the association was also shown in test set 2 ($p = 0.054$; Figure 3(e)).

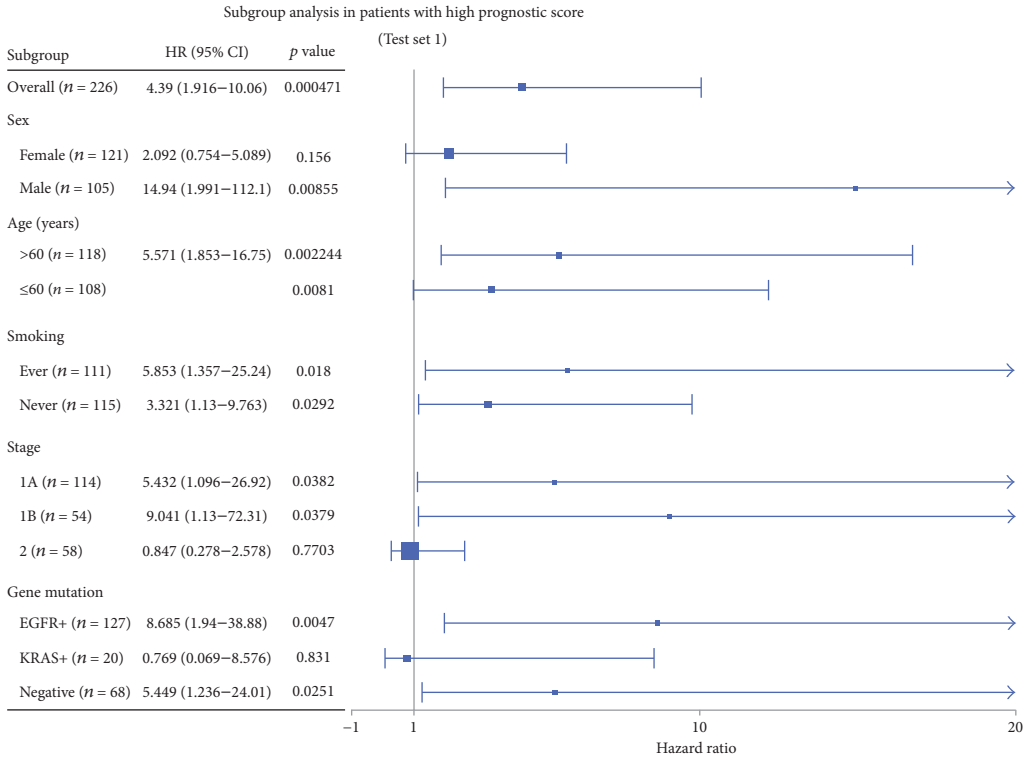
3.4. NetScore as an Independent Predictive Factor for Prognosis. Cox multivariate analysis revealed that the risk group was associated with OS independent of stage as well as other clinicopathological features in the training set and test set 1 (Table 1). The independent predictive factors for OS in Cox multivariate analysis were the risk group ($p = 0.001$) and T-stage 3 ($p = 0.030$) in training set and the risk group ($p = 0.01$) and EGFR mutation status ($p = 0.005$) in test set 1. In test set 2, there was no feature significantly associated with OS in univariate Cox analysis, though the high-risk group showed a trend of unfavorable prognosis ($p = 0.06$). We also evaluated the prognostic value of NetScore in subgroups divided by clinical and pathological features. In the training set, the high-risk group was significantly associated with poor prognosis in subgroups regardless of age and T-stage. In all subgroups, a trend of close relationship between

Table 1: Univariate and multivariate Cox regression analysis of the risk stratification model and clinicopathological variables.

Variables	Univariate analysis		Multivariate analysis	
	Hazard ratio (95% CI)	p value	Hazard ratio (95% CI)	p value
Training set				
Gene network prognostic score, high risk group	2.51 (1.65–3.81)	<0.001	3.057 (1.556–6.006)	0.001
Age, older than 60	1.66 (0.96–2.86)	0.068		
Sex, male	1.28 (0.86–1.88)	0.221		
Smoking status, ex-smoker	0.59 (0.31–1.12)	0.108		
Smoking status, never smoker	0.51 (0.22–1.19)	0.120		
T stage: II	2.50 (1.31–4.79)	0.006	1.266 (0.599–2.674)	0.537
T stage: III	13.32 (2.89–61.32)	0.001	5.895 (1.189–29.237)	0.030
N stage: I	2.27 (1.28–4.05)	0.005	1.762 (0.943–3.294)	0.076
Test set 1				
Gene network prognostic score, high risk group	4.39 (1.92–10.06)	0.0004	2.97 (1.25–7.09)	0.01
Age, older than 60	1.27 (0.65–2.48)	0.49		
Sex, male	1.52 (0.78–2.96)	0.22		
Smoking status, never smoker	0.61 (0.31–1.19)	0.15		
Stage: II	4.23 (2.17–8.24)	0.00002		
EGFR mutation +	0.47 (0.24–0.93)	0.03	2.74 (1.36–5.54)	0.005
KRAS mutation +	0.87 (0.27–2.85)	0.82	0.64 (0.32–1.27)	0.20
Test set 2				
Gene network prognostic score, high risk group	1.81 (0.98–3.36)	0.06		
Age, older than 60	1.33 (0.73–2.43)	0.35		
Sex, male	0.83 (0.40–1.75)	0.63		
Stage: T2	1.65 (0.84–3.25)	0.14		



(a)



(b)

Figure 4: Continued.

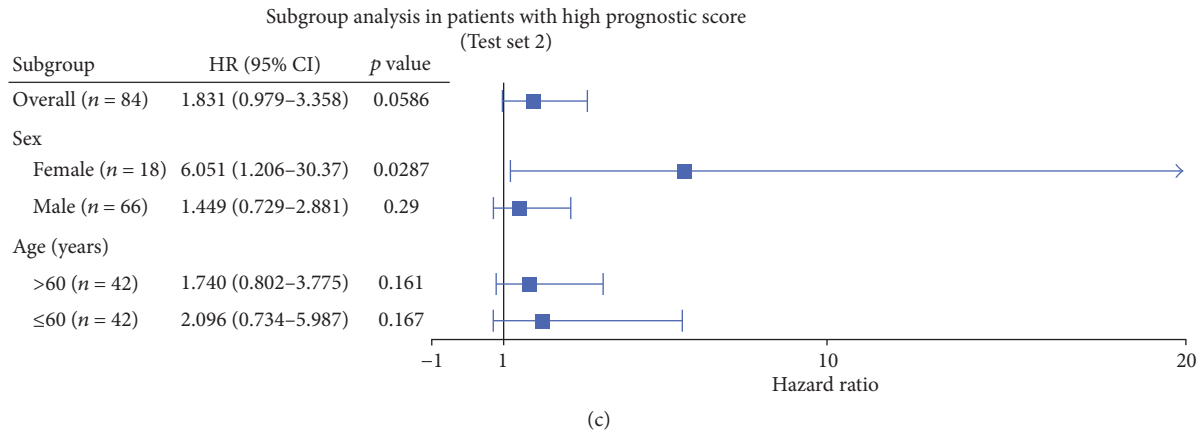


FIGURE 4: *Subgroup analysis using NetScore*. (a) Predictive value of our risk stratification model was tested in subgroups classified by clinicopathological characteristics of the training set. A trend of association between the risk group and overall survival was found in all subgroups. (b, c) The same subgroup survival analysis was also performed in both test sets. (b) The risk group was associated with overall survival regardless of clinicopathological variables except female, stage II, and KRAS mutation subgroups in test set 1. (c) Regardless of subgroups, a trend of poor prognosis in high-risk group was also found in test set 2.

the risk group and OS was found except never-smoking subgroup (Figure 4(a), Supplementary Figure 4). According to subgroup analysis in test set 1, the risk group was closely associated with OS in male, old-aged, ever/never smokers, stage IA/IB, EGFR positive and all negative mutation subgroups (Figure 4(b), Supplementary Figure 5). A trend of association between the risk group and OS was also revealed in each subgroup of test set 2, regardless of clinical features including sex, age, and T-stage (Figure 4(c), Supplementary Figure 6).

4. Discussion

In this study, we developed a risk stratification model for lung adenocarcinoma based on gene coexpression networks and deep learning. Survival-related network modules were identified in multiple cohorts and representative genes of these modules were selected for risk stratification modeling. The model constructed by deep CNN reflects gene expression patterns of survival-related network modules and it provides prognostic score, NetScore. The NetScore was significantly associated with OS in all cohorts and also an independent predictor for OS from clinicopathological variables.

The model based on survival-related network modules can provide more robust risk stratification compared with models focusing on statistical combination of individual prognostic genes which have been proposed in the previous studies [1–6]. In spite of previous promising results of individual gene-based models, they failed to validate in independent samples of other study [4]. Furthermore, there were few overlapping significant prognostic genes in the previous models. A meta-analysis of published gene expression data revealed that few genes were associated with survival of lung adenocarcinoma [15]. The result of few significant prognostic genes in large samples implied the limitation of usage of individual genes for risk stratification. Besides, selection of individual significant genes has a substantial problem of multiple statistical testing [16]. Instead of these previous approaches, systemic approach integrating gene interaction as well as individual genes would be a breakthrough for robust risk stratification modeling because variation patterns of their expression levels can be associated with prognosis.

Recently, DL has dramatically improved data analysis in genomics and imaging fields [17, 18]. The main contribution of DL for our risk stratification model is to apply deep neural network to gene expression data. It employed convolutional layers for extracting multiple gene expression patterns. Another contribution is to solve regression problems of survival data by using a specialized loss function [19] (see Supplementary Methods). We compared predictive accuracy of DL-based model and conventional Cox proportional hazard model obtained from the expression level of selected 20 genes. Predictability of the DL-based model was significantly higher than that of the Cox model in test set 1 (C-index = 0.709 ± 0.042 and 0.608 ± 0.046, resp.; $p = 0.004$). It was also higher in the training set and test set 2 though the difference did not reach statistical significance (Supplementary Methods, Supplementary Figure 7). Furthermore, to confirm robustness of NetScore, the model was retrained by the dataset combined by original training and test set 1 and validated in test set 2. NetScore of the retrained model was also significantly associated with OS in the test set 2 ($p = 0.003$; C-index = 0.651 ± 0.042). To our knowledge, NetScore is the first study that apply deep convolutional neural network to high-dimensional gene expression data for predicting prognosis. By applying this novel approach to various genomic data, risk stratification and survival prediction could be improved compared with conventional Cox model.

NetScore was trained by various samples with different clinicopathological characteristics. We found NetScore was associated with sex, smoking status, stage, and molecular

subtypes (Supplementary Figure 8). Briefly, a trend of high NetScore was found in male, smokers, late stage, and KRAS mutation positive samples. Nonetheless, NetScore was significantly associated with OS independent of clinicopathological variables according to multivariate and subgroup analyses. Of note, NetScore was significant predictor in early stage subgroups (stage IA/IB). This finding could be important because the new risk stratification could identify patients who might need adjuvant chemotherapy. For example, a recent clinical trial using 15-gene signature based on individual prognostic genes showed successful selection of patients with stage IB and II NSCLC who would most likely benefit from adjuvant chemotherapy [20]. In the future, as a new prognostic biomarker based on gene network, the usefulness of NetScore should be tested whether it could affect clinical decision and compared with the previous prognostic models using individual genes.

We developed a risk stratification model for lung adenocarcinoma using gene coexpression network. A future extension of our work would be to apply this approach to the coexpression networks of other cancer types. In terms of technical improvement, modification of DL architecture and selection process of representative genes could improve the prediction accuracy. Finally, we expected that a prospectively designed clinical trial with well-controlled clinicopathological variables would help find clinical application of our new risk stratification model.

Disclosure

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of Interest

The authors declare no financial conflicts of interest.

Authors' Contributions

Hongyoon Choi and Kwon Joong Na contributed equally to this work.

Supplementary Materials

Supplementary Methods for public data collection and preprocessing, functional annotation and network visualization of survival-related network modules, representative genes selection for risk stratification model construction, comparison of predictability between DL-based model and conventional Cox proportional hazard model, and convolutional neural network for risk stratification. Supplementary Figure 1: gene coexpression network topology analysis for various soft threshold powers. Supplementary Figure 2: the relationship between coexpression network modules. Supplementary Figure 3: representative genes selected from the red and turquoise modules. Supplementary Figure 4: subgroup analysis using NetScore in the training set. Supplementary Figure 5: subgroup analysis using NetScore in the test set 1. Supplementary Figure 6: subgroup analysis using NetScore in the test set 2. Supplementary Figure 7: comparison of predictability between deep neural network and conventional Cox regression models. Supplementary Figure 8: association of gene network prognostic score (NetScore) with clinicopathological variables. Supplementary Table 1: microarray data sets from Gene Expression Omnibus used in this study. Supplementary Table 2: demographic and baseline clinical characteristics of patients. Supplementary Table 3: significantly enriched gene ontology biological process terms of five survival-related network modules. *(Supplementary Materials)*

References

[1] H. Y. Chen, S. L. Yu, C. H. Chen et al., "A five-gene signature and clinical outcome in non-small-cell lung cancer," *The New England Journal of Medicine*, vol. 356, no. 1, pp. 11–20, 2007.

[2] Y. Xie, G. Xiao, K. R. Coombes et al., "Robust gene expression signature from formalin-fixed paraffin-embedded samples predicts prognosis of non-small-cell lung cancer patients," *Clinical Cancer Research*, vol. 17, no. 17, pp. 5705–5714, 2011.

[3] M. Skrzypski, E. Jassem, M. Taron et al., "Three-gene expression signature predicts survival in early-stage squamous cell carcinoma of the lung," *Clinical Cancer Research*, vol. 14, no. 15, pp. 4794–4799, 2008.

[4] K. Shedden, J. M. G. Taylor, S. A. Enkemann et al., "Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study," *Nature Medicine*, vol. 14, no. 8, pp. 822–827, 2008.

[5] P. Roepman, J. Jassem, E. F. Smit et al., "An immune response enriched 72-gene prognostic profile for early-stage non-small-cell lung cancer," *Clinical Cancer Research*, vol. 15, no. 1, pp. 284–290, 2009.

[6] P. C. Boutros, S. K. Lau, M. Pintilie et al., "Prognostic gene signatures for non-small-cell lung cancer," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 106, no. 8, pp. 2824–2828, 2009.

[7] A. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature Reviews Genetics*, vol. 12, no. 1, pp. 56–68, 2011.

[8] B. Zhang and S. Horvath, "A general framework for weighted gene co-expression network analysis," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, no. 1, article 1128, 2005.

[9] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2015," *CA: A Cancer Journal for Clinicians*, vol. 65, no. 1, pp. 5–29, 2015.

[10] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, and J. Lortet-Tieulent, "Global cancer statistics, 2012," *CA: A Cancer Journal for Clinicians*, vol. 65, no. 2, pp. 87–108, 2015.

[11] W. Pao and N. Girard, "New driver mutations in non-small-cell lung cancer," *The Lancet Oncology*, vol. 12, no. 2, pp. 175–180, 2011.

[12] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, article 559, 2008.

[13] P. Langfelder, B. Zhang, and S. Horvath, "Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for R," *Bioinformatics*, vol. 24, no. 5, pp. 719-720, 2008.

[14] F. E. Harrell, K. L. Lee, R. M. Califf, D. B. Pryor, and R. A. Rosati, "Regression modelling strategies for improved prognostic prediction," *Statistics in Medicine*, vol. 3, no. 2, pp. 143–152, 1984.

[15] B. Győrffy, P. Surowiak, J. Budczies, and A. Lánczky, "Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in non-small-cell lung cancer," *PLoS ONE*, vol. 8, no. 12, Article ID e82241, 2013.

[16] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour, "Microarray data analysis: from disarray to consolidation and consensus," *Nature Reviews Genetics*, vol. 7, no. 1, pp. 55–65, 2006.

[17] C. Angermueller, T. Pärnamaa, L. Parts, and O. Stegle, "Deep learning for computational biology," *Molecular Systems Biology*, vol. 12, no. 7, article no. 878, 2016.

[18] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[19] J. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "DeepSurv: personalized treatment recommender system using a cox proportional hazards deep neural network," https://arxiv.org/abs/1606.00931.

[20] C.-Q. Zhu, K. Ding, D. Strumpf et al., "Prognostic and predictive gene signature for adjuvant chemotherapy in resected non-small-cell lung cancer," *Journal of Clinical Oncology*, vol. 28, no. 29, pp. 4417–4424, 2010.

Hindawi
BioMed Research International
Volume 2018, Article ID 6916797, 7 pages
https://doi.org/10.1155/2018/6916797

Research Article

Glycyrrhizin Suppresses the Growth of Human NSCLC Cell Line HCC827 by Downregulating HMGB1 Level

Xiaojin Wu,[1] Weitao Wang,[2] Yuanyuan Chen,[1] Xiangqun Liu,[3] Jindong Wang,[4] Xiaobin Qin,[5] Dawei Yuan,[2] Tao Yu,[5] Guangxia Chen,[6] Yanyan Mi,[7] Jie Mou,[7] Jinpeng Cui,[8] Ankang Hu,[9] Yunxiang E,[5] and Dongsheng Pei[10]

[1] *Department of Radiation Oncology, The First People's Hospital of Xuzhou, Xuzhou, Jiangsu 221002, China*
[2] *Genesis Beijing Co., Ltd., Beijing 100102, China*
[3] *Department of Respiratory Diseases, The First People's Hospital of Xuzhou, Xuzhou, Jiangsu 221002, China*
[4] *Department of Chest Surgery, The First People's Hospital of Xuzhou, Xuzhou, Jiangsu 221002, China*
[5] *Department of Tumor, The First People's Hospital of Xuzhou, Xuzhou, Jiangsu 221002, China*
[6] *Department of Gastroenterology, The First People's Hospital of Xuzhou, Xuzhou, Jiangsu 221002, China*
[7] *Department of Pharmacy, Xuzhou Medical College, Xuzhou, Jiangsu 221004, China*
[8] *Clinical Laboratory of Yantaishan Hospital, No. 91, Jiefang Road, Yantai, Shandong 264001, China*
[9] *Laboratory Animal Center, Xuzhou Medical College, Xuzhou, Jiangsu 221004, China*
[10] *Department of Pathology, Xuzhou Medical College, Xuzhou, Jiangsu 221004, China*

Correspondence should be addressed to Dongsheng Pei; dspei@xzhmu.edu.cn

Received 9 October 2017; Revised 24 November 2017; Accepted 29 November 2017; Published 15 January 2018

Academic Editor: Jialiang Yang

Lung cancer has very high mortality and glycyrrhizin was found to significantly inhibit the growth of lung cancer cells *in vitro* and tissues in mice. However, the detailed inhibitory role of glycyrrhizin in the growth of lung cancer is still unclear. In this study, we first found that glycyrrhizin inhibited the growth of lung tumor in PDX mice. And high level of HMGB1 promoted the migration and invasion of lung cancer cells, which was suppressed by glycyrrhizin. Moreover, glycyrrhizin reduced the activity of JAK/STAT signaling pathway, which is the upstream regulator of HMGB1. Therefore, this study revealed a potential mechanism by which glycyrrhizin can inhibit the progression of lung cancer.

1. Introduction

Lung cancer is highly prevalent worldwide. An estimated 526,510 men and women living in the United States had a history of lung cancer in 2016 [1]. Traditionally, lung cancer can be divided into two types: small cell lung cancer (10–15%) and non-small cell lung cancer (NSCLC) (85–90%). The 5-year survival for NSCLC (21%) is usually higher than small cell lung cancer (7%). However, the 5-year relative survival rate of lung cancer is generally lower (17%) than other cancer types [1, 2]. Therefore, it is necessary to design novel drugs for effective treatment of lung cancer.

Glycyrrhizin, a glycoconjugated triterpene, is extracted from the roots of licorice plant, *Glycyrrhiza glabra*. It was first identified as an antiviral drug [3] and has been subsequently used in the treatment of patients with chronic hepatitis B and hepatitis C due to its anti-inflammatory role [4, 5]. Recently, glycyrrhizin was found to suppress lung adenocarcinoma A549 cell growth by inducing cancer cell apoptosis through downregulating the activity of thromboxane synthase pathway [6]. Furthermore, the growth of lung tumor tissue in PDX mouse model could be effectively inhibited by combining glycyrrhizin with cisplatin treatment, which showed low toxicity and side effects [6, 7]. Therefore, glycyrrhizin could be developed as a drug for lung cancer therapy.

High Mobility Group Box 1 (HMGB1) is a conserved nonhistone chromosomal protein that regulates nucleosome formation and gene transcription. HMGB1 could also be

released into extracellular matrix, where it is recognized as a cytokine based on its role in mediating systemic inflammatory response [8]. Further, HMGB1 is involved in cancer progression and observed in various types of cancers [9–11]. High level of HMGB1 is always associated with cancer metastasis, suggesting that HMGB1 promotes invasion and metastasis of cancer cells [12–16]. Similarly, HMGB1 can promote the migration and invasion of lung cancer cells and regulate the metastasis of lung cancer [17, 18]. High serum level of HMGB1 can be a potential clinical biomarker for lung cancer [19–22].

Glycyrrhizin was found to inhibit HMGB1 for diagnosing HMGB1-mediated lethal systemic inflammation [23]. Glycyrrhizin could block the chemoattractant and mitogenic activities of HMGB1 by directly binding to its two HMG-box [24]. Moreover, glycyrrhizin could also regulate the expression of HMGB1 after hepatic ischemia-reperfusion (I/R) injury [25] and induction of traumatic pancreatitis in rats [26]. Moreover, glycyrrhizin can suppress tumor growth by reducing the level or activity of HMGB1 [27, 28]. However, whether the anti-lung cancer effect of glycyrrhizin involves suppression of HMGB1 remains unknown.

In this study, glycyrrhizin was found to suppress the tumor growth of NSCLC in PDX mice. Furthermore, the levels of both HMGB1 and its related JAK/STAT3 signal pathway factors were downregulated by treating PDX mice with glycyrrhizin. These findings indicated that glycyrrhizin may be involved in anticancer therapy of NSCLC through downregulating the level of HMGB1.

2. Materials and Methods

2.1. Cell Lines. Human NSCLC cell line HCC827 was purchased from the cell bank of Chinese Academy of Sciences and cultured in RPMI 1640 medium containing 10% fetal bovine serum (FBS).

2.2. PDX Mice Model. The logarithmic growth phase of HCC827 cells was suspended in cell culture medium and the concentration was adjusted to 1×10^7 ml^{-1}. Then, 0.2 ml cell suspension was subcutaneously injected into the back of nude mice, and the tumor formation was visually observed after 5 d. When the tumor grew to a diameter of 5–7 mm, 30 mice were randomly divided into the model group and glycyrrhizic acid treatment group, with 15 rats/group.

2.3. Glycyrrhizin Treatment. Glycyrrhizic acid (purchased from Dalian Meilen Pharmaceutical Technology Development Co., Ltd.) was diluted in DMSO and intraperitoneally injected at a dose of 100 mg/kg for two days, followed by continuous administration for two weeks. The model control group was injected with the same amount of DMSO.

2.4. Hematoxylin and Eosin Staining. Lung cancer cells from the glycyrrhizin treated and DMSO treated groups were stained by Hematoxylin and Eosin (H&E), as previously described [29]. First, samples were treated with distilled water and the nuclei were stained with the alum haematoxylin. Then they differentiated with 0.3% acid alcohol and were stained with eosin. Finally, they were dehydrated through 95% alcohol, cleared in 2 changes of xylene, and mounted with xylene based mounting medium.

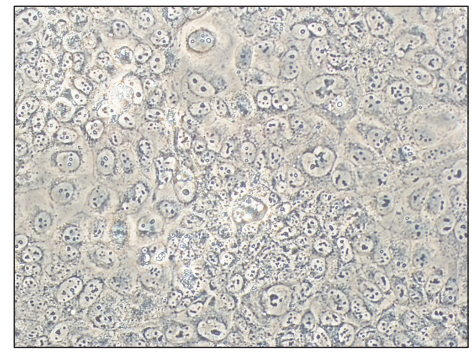


Figure 1: Morphology of lung cancer HCC827 cells in PDX mice.

2.5. Immunohistochemical Staining. The immunohistochemical staining was performed using EnvisionTM, as previously described [30]. Lung cancer tissue staining results were evaluated by IRS score under double-blind conditions by two senior pathological experts. The average staining score was calculated by combining the positive staining intensity and the percentage of positive cells.

2.6. Western Blot Analysis. Western blot analysis was performed as previously described [31]. Cells were lysed using RIPA lysis buffer and then proteins were extracted. Equal amounts of proteins were denatured in boiling SDS sample buffer and subjected to 10% SDS-PAGE. Then, the proteins were transblotted onto polyvinylidene difluoride membranes with a wet blot system. The membranes were blocked by 5% dry skim milk and incubated with primary anti-HMGB1 (Abcam), anti-P-Jak2, P-Stat3, Jak2, and Stat3 (Cell Signaling) antibodies. β-Actin was used as an internal control. Finally, the membranes were treated with enhanced chemiluminescent system for visualization of the protein bands. The bands were quantified using Image J software.

3. Results

3.1. Establishment of the PDX Mouse Model for Non-Small Cell Lung Cancer. For testing the effect of glycyrrhizin on the growth of NSCLC, we established the PDX mouse model using the lung cancer HCC827 cell line. Then, we observed the morphology of the cancer cells in PDX mice and found that the arrangement of cells and shape of nuclei were irregular (Figure 1).

3.2. Glycyrrhizin Inhibits Tumor Growth in the PDX Mouse Model. We detected the effect of glycyrrhizin on tumor growth in PDX mice. The PDX mice were divided into two groups: glycyrrhizin treated and DMSO treated. We first observed the morphology of cancer cells by HE staining (Figure 2(a)). Next, we evaluated the tumor sizes and weights in the two groups. The glycyrrhizin treated mice had smaller

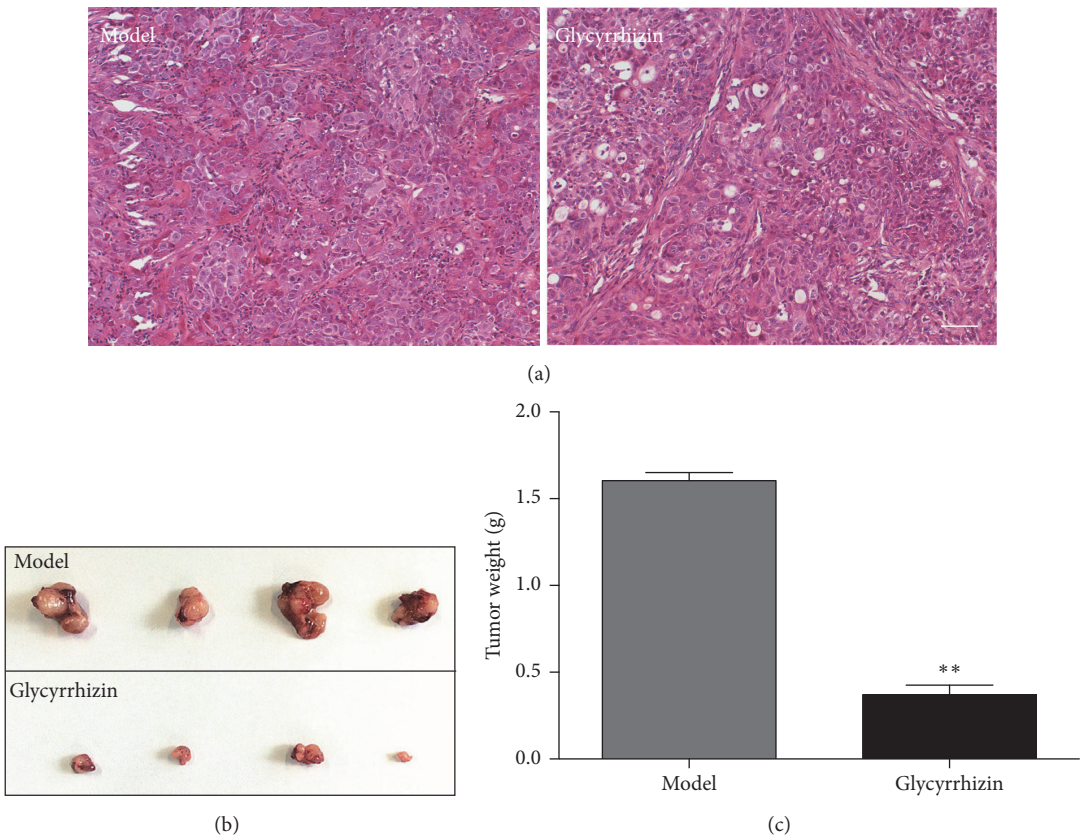


Figure 2: Glycyrrhizin inhibits tumor growth in PDX mice. (a) indicates the morphology of lung cancer cells from the two groups by HE staining, bar = 100 μm. (b) indicates the size of tumors from the two groups. Tumors were excised from the two groups and the tumor sizes between the two groups were compared. (c) indicates the average weight of tumors in the two groups. The difference of tumor weights between the glycyrrhizin treatment and DMSO treatment (model) was significant. $^{**}p < 0.01$, as compared to the model control by t-test.

tumor sizes than the DMSO treated mice (Figure 2(b)). The average tumor weight of the glycyrrhizin treated mice was significantly lower than the DMSO treated group ($p < 0.01$) (Figure 2(c)). These results suggested that glycyrrhizin can inhibit the growth of lung tumor in PDX mice.

3.3. HMGB1 Protein Is Suppressed by Glycyrrhizin. HMGB1 was reported to promote the migration and invasion of lung cancer cells and facilitate lung cancer metastasis [17, 18]. Glycyrrhizin functions as an inhibitor of HMGB1 by blocking its activity [24] or downregulating its expression [25]. However, whether the anticancer effect of glycyrrhizin in lung cancer relies on downregulating the expression of HMGB1 is unclear. So we detected the protein expression of HMGB1 in lung cancer tissues obtained from the two groups by IHC staining, and representative images are shown in Figure 3(a). Furthermore, the protein level of HMGB1 in the glycyrrhizin treated group was significantly lower than that in the DMSO treated group ($p < 0.05$) (Figure 3(b)). We also detected the protein level of HMGB1 by western blot and observed similar results. Interestingly, higher level of HMGB1 was also seen in lung tumor tissue from PDX mice compared to that from normal mice (NC), which is consistent with previous reports indicating that HMGB1 is related to cancer progression. However, the level of HMGB1 obviously decreased after glycyrrhizin treatment (Figure 3(c)), suggesting that HMGB1 protein is suppressed by glycyrrhizin.

3.4. Glycyrrhizin Inhibits the Phosphorylation of Jak2 and Stat3. In mammals, the JAK/STAT pathway is the principal signaling mechanism of inflammation and involves various cytokines and growth factors [32, 33]. Members of the JAK family are receptor-associated tyrosine kinases activated by various extracellular signals. Signal Transducer and Activator of Transcription (STAT) proteins are the typical substrates of JAK kinases and are generally associated with transcriptional activation as transcription factors [34]. HMGB1 is released from the nucleus into the cytoplasm, which is regulated by JAK/STAT signal pathway mediated HMGB1 hyperacetylation [35]. Furthermore, resveratrol could reduce the release of HMGB1 from the nucleus to the cytoplasm by suppressing the activity of STAT signaling pathway [36].

We further examined whether glycyrrhizin could inhibit the activity of JAK/STAT signaling pathway. The phosphorylation status of Jak2 and Stat3 was detected by specific phosphorylated antibodies, which showed that the phosphorylation levels of Jak2 and Stat3 were significantly higher in the PDX-model mice, but obviously lower after glycyrrhizin

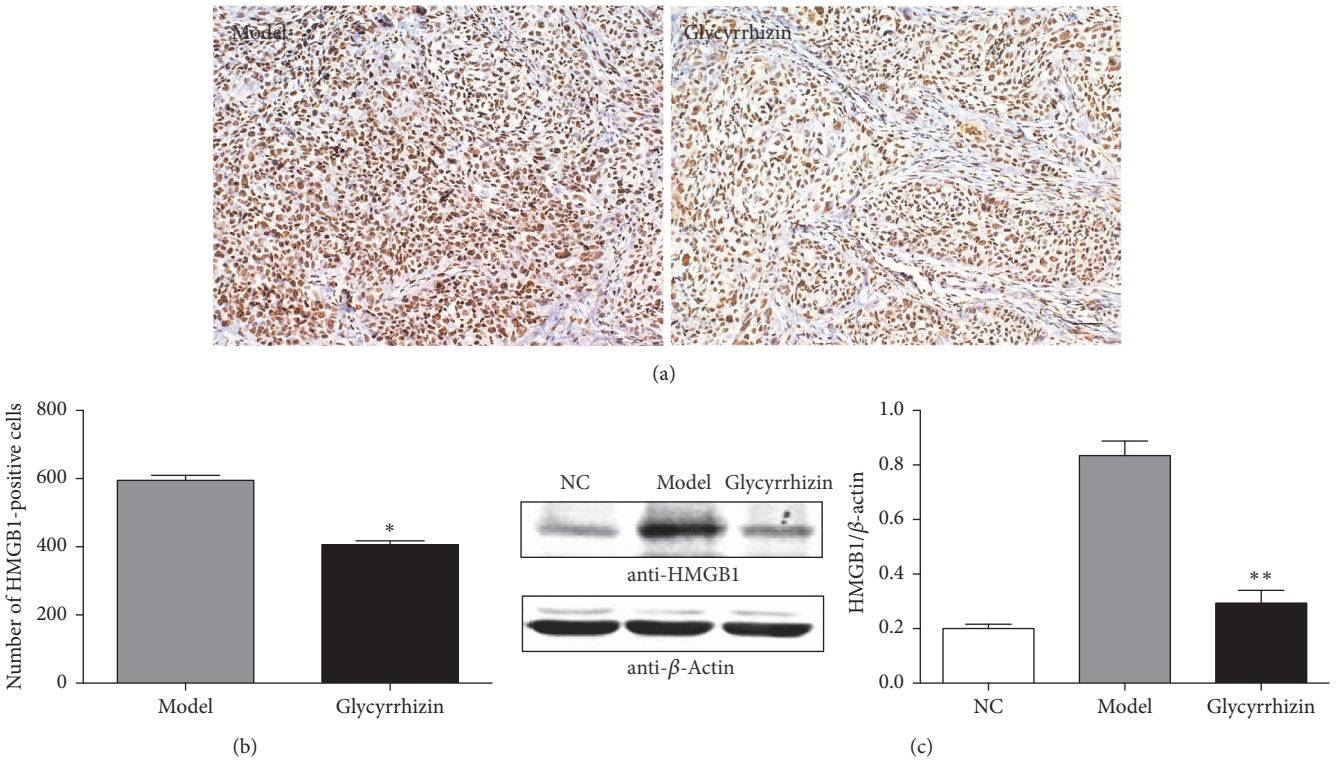


Figure 3: The protein level of HMGB1 is suppressed by glycyrrhizin. (a) indicates the protein level of HMGB1 in lung tumor tissues from the two groups detected by IHC staining assay, bar = 100 μm. (b) indicates the quantitative results based on IHC staining assay. HMGB1-positive cells were counted in the two groups, respectively. $^{*}p < 0.05$, as compared to the model control by t-test. (c) indicates the protein level of HMGB1 from the three groups by western blot. β-Actin was used as an internal control. p value was calculated by t-test between glycyrrhizin treatment and DMSO treatment (model). "NC" represents lung tissue from normal mice. $^{**}p < 0.01$, as compared to the model control.

treatment (Figures 4(a) and 4(b)). These results indicated that glycyrrhizin can inhibit the activity of JAK/STAT signaling pathway, which is the upstream regulator of HMGB1.

4. Discussion

In this study, glycyrrhizin was shown to suppress the growth of lung tumor tissues in PDX mice, derived from NSCLC HCC827 cell line, which is consistent with recent reports on the anticancer effect of glycyrrhizin on lung cancer progression [6, 7]. Huang et al. [6] showed that glycyrrhizin could inhibit the lung adenocarcinoma A549 cell line growth both *in vitro* and in PDX mice. Subsequently, Deng et al. [7] reported that glycyrrhizin combined with cisplatin had a better anticancer effect in the PDX mice model. We used another NSCLC cell line for establishing the PDX mice and proved that anticancer effect of glycyrrhizin was similar. Our findings further confirmed that glycyrrhizin may be a potential anticancer drug for NSCLC.

HMGB1, a cytokine, has extracellular functions in inflammation and cancer progression. As a late modulator of inflammation, HMGB1 functions as a damage-associated molecular pattern in the sterile inflammation model by amplifying hepatic ischemia/reperfusion (I/R) and acetaminophen-induced liver necrotic injury [37, 38]. Glycyrrhizin can antagonize the inflammatory effect of HMGB1 by suppressing the expression of HMGB1 in hepatic I/R injury [25]. In addition, HMGB1 has a critical role in cancer metastasis. High levels of HMGB1 are always observed in various cancer types, including ovarian [15], liver [37], and lung [22] cancers. HMGB1 can promote the migration and invasion of lung cancer cells [18]. Drugs designed for regulating the level of HMGB1 may have a potential clinical value for lung cancer patients. In this study, high level of HMGB1 was related to the growth of lung tumors in PDX mice. Glycyrrhizin obviously inhibited the level of HMGB1. Therefore, glycyrrhizin acts as an inhibitor of HMGB1 and the growth of lung tumor. The direct role of HMGB1 in anticancer effect of glycyrrhizin on lung cancer progression needs further study.

Glycyrrhizin influenced the upstream regulator of HMGB1, the JAK/STAT signaling pathway. HMGB1 release from nucleus to cytoplasm is regulated by the activation of JAK/STAT signaling pathway [35]. We found that glycyrrhizin can block the activity of JAK/STAT signaling pathway by inhibiting phosphorylation of Jak2 and Stat3. This is the first study to show that the level of HMGB1 may be controlled by the JAK/STAT signaling pathway.

5. Conclusion

Glycyrrhizin inhibits the growth of lung tumors in PDX mice by downregulating the level of HMGB1. This mechanism is

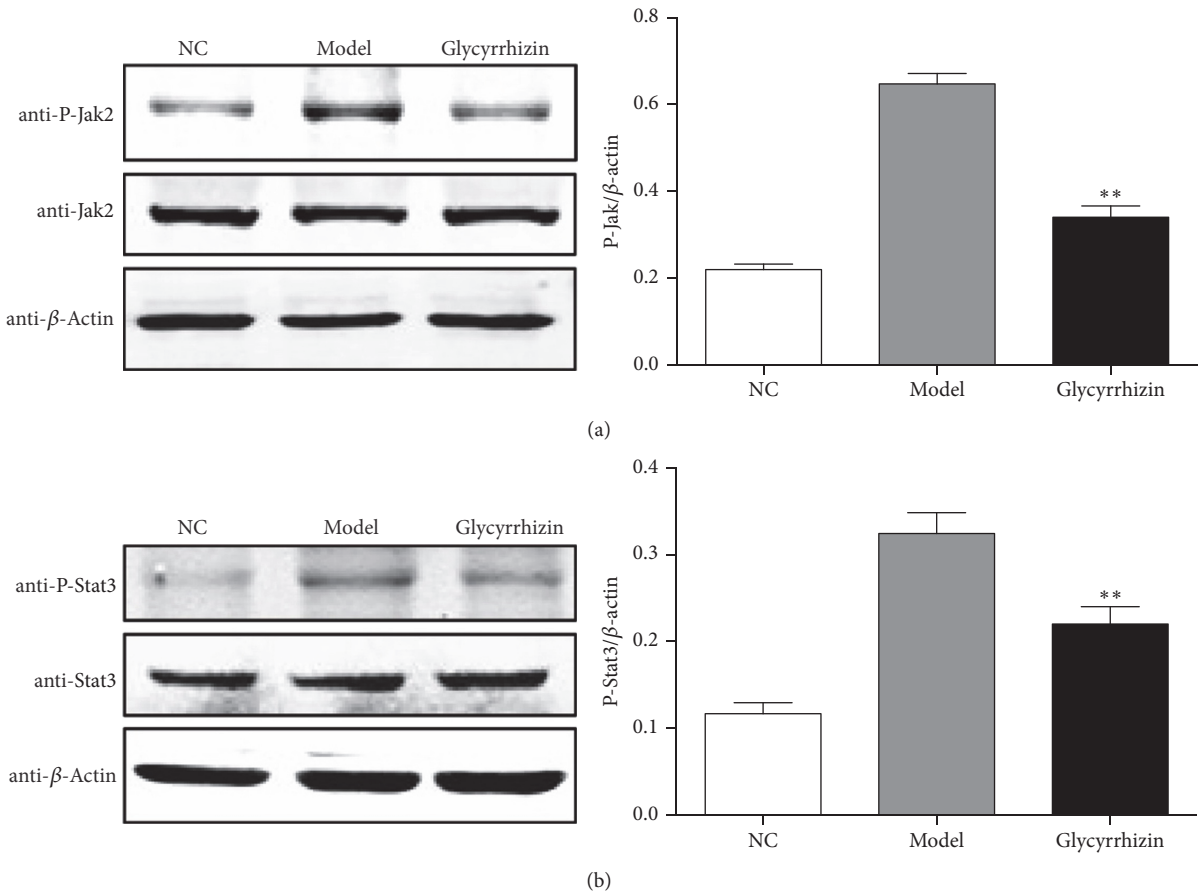


Figure 4: Glycyrrhizin inhibits the phosphorylation of Jak2 and Stat3. The phosphorylation level of Jak2 was detected by western blot in the three experimental groups, respectively. $^{**}p < 0.01$, as compared to the model control (a). The phosphorylation level of Stat3 was detected by western blot in the three experimental groups, respectively. $^{**}p < 0.01$, as compared to the model control (b). β-Actin was used as an internal control. p value was calculated by t-test between glycyrrhizin treatment and DMSO treatment (model). "NC" represents lung tissue from normal mice. Stars are showing p values between model and glycyrrhizin.

potentially due to the inhibition of JAK/STAT signaling pathway by glycyrrhizin. Glycyrrhizin may be further investigated as a potential drug for NSCLC.

Disclosure

Xiaojin Wu and Weitao Wang are co-first authors.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This study was supported by the Foundation of Jiangsu Provincial Commission of Health and Family Planning (Grant QMRC2016363), Key Talents of Medical Science in Jiangsu Province (Grant Z201627), Xuzhou Administration of Science & Technology (Grant KC14SH009), and the 333 High-Level Talents of Jiangsu Province (Grant (2016)-1903).

References

[1] K. D. Miller, R. L. Siegel, and C. C. Lin, "Cancer treatment and survivorship statistics, 2016," *CA: A Cancer Journal for Clinicians*, vol. 66, no. 4, pp. 271–289, 2016.

[2] S. Subramaniam, R. K. Thakur, V. K. Yadav, R. Nanda, S. Chowdhury, and A. Agrawal, "Lung cancer biomarkers: state of the art," *Journal of Carcinogenesis*, vol. 12, no. 3, 2013.

[3] R. Pompei, O. Flore, M. A. Marccialis, A. Pani, and B. Loddo, "Glycyrrhizic acid inhibits virus growth and inactivates virus particles," *Nature*, vol. 281, no. 5733, pp. 689-690, 1979.

[4] S. Iino, T. Tango, T. Matsushima et al., "Therapeutic effects of stronger neo-minophagen C at different doses on chronic hepatitis and liver cirrhosis," *Hepatology Research*, vol. 19, no. 1, pp. 31–40, 2001.

[5] K. Miyake, T. Tango, Y. Ota et al., "Efficacy of Stronger Neo-Minophagen C compared between two doses administered three times a week on patients with chronic viral hepatitis," *Journal of Gastroenterology and Hepatology*, vol. 17, no. 11, pp. 1198–1204, 2002.

[6] R.-Y. Huang, Y.-L. Chu, Z.-B. Jiang, X.-M. Chen, X. Zhang, and X. Zeng, "Glycyrrhizin suppresses lung adenocarcinoma cell growth through inhibition of thromboxane synthase," *Cellular Physiology and Biochemistry*, vol. 33, no. 2, pp. 375–388, 2014.

[7] Q.-P. Deng, M.-J. Wang, X. Zeng, G. G. Chen, and R.-Y. Huang, "Effects of glycyrrhizin in a mouse model of lung adenocarcinoma," *Cellular Physiology and Biochemistry. International Journal of Experimental Cellular Physiology, Biochemistry and Pharmacology*, vol. 41, Article ID 13831392, pp. 1383–1392, 2017.

[8] M. T. Lotze and K. J. Tracey, "High-mobility group box 1 protein (HMGB1): nuclear weapon in the immune arsenal," *Nature Reviews Immunology*, vol. 5, no. 4, pp. 331–342, 2005.

[9] D. Süren, M. Yildirim, Ö. Demirpençe et al., "The role of High Mobility Group Box 1 (HMGB1) in colorectal cancer," *Medical Science Monitor*, vol. 20, pp. 530–537, 2014.

[10] Y. R. Choi, H. Kim, H. J. Kang et al., "Overexpression of high mobility group box 1 in gastrointestinal stromal tumors with KIT mutation," *Cancer Research*, vol. 63, Article ID 21882193, pp. 2188–2193, 2003.

[11] Y. Li, J. Tian, X. Fu et al., "Serum high mobility group box protein 1 as a clinical marker for ovarian cancer," *Neoplasma*, vol. 61, no. 5, pp. 579–584, 2014.

[12] W. Yan, Y. Chang, X. Liang et al., "High-mobility group box 1 activates caspase-1 and promotes hepatocellular carcinoma invasiveness and metastases," *Hepatology*, vol. 55, no. 6, pp. 1863–1875, 2012.

[13] B. Song, W.-G. Song, Z.-J. Li et al., "Effect of HMGB1 silencing on cell proliferation, invasion and apoptosis of MGC-803 gastric cancer cells," *Cell Biochemistry & Function*, vol. 30, no. 1, pp. 11–17, 2012.

[14] S. Jube, Z. S. Rivera, M. E. Bianchi et al., "Cancer cell secretion of the DAMP protein HMGB1 supports progression in malignant mesothelioma," *Cancer Research*, vol. 72, no. 13, pp. 3290–3301, 2012.

[15] J. Chen, X. Liu, J. Zhang, and Y. Zhao, "Targeting HMGB1 inhibits ovarian cancer growth and metastasis by lentivirus-mediated RNA interference," *Journal of Cellular Physiology*, vol. 227, no. 11, pp. 3629–3638, 2012.

[16] C. A. Wild, S. Brandau, R. Lotfi et al., "HMGB1 is overexpressed in tumor cells and promotes activity of regulatory T cells in patients with head and neck cancer," *Oral Oncology*, vol. 48, no. 5, pp. 409–416, 2012.

[17] C. Zhang, S. Ge, C. Hu, N. Yang, and J. Zhang, "MiRNA-218, a new regulator of HMGB1, suppresses cell migration and invasion in non-small cell lung cancer," *Acta Biochimica et Biophysica Sinica*, vol. 45, Article ID 10551061, pp. 1055–1061, 2013.

[18] C. Wang, G. Fei, Z. Liu, Q. Li, Z. Xu, and T. Ren, "HMGB1 was a pivotal synergistic effecor for CpG oligonucleotide to enhance the progression of human lung cancer cells," *Cancer Biology & Therapy*, vol. 13, no. 9, pp. 727–736, 2012.

[19] G.-H. Shang, C.-Q. Jia, H. Tian et al., "Serum high mobility group box protein 1 as a clinical marker for non-small cell lung cancer," *Respiratory Medicine*, vol. 103, no. 12, pp. 1949–1953, 2009.

[20] F. Wei, F. Yang, X. Jiang, W. Yu, and X. Ren, "High-mobility group nucleosome-binding protein 1 is a novel clinical biomarker in non-small cell lung cancer," *Tumor Biology*, vol. 36, no. 12, pp. 9405–9410, 2015.

[21] A. Feng, Z. Tu, and B. Yin, "The effect of HMGB1 on the clinicopathological and prognostic features of non-small cell lung cancer," *Oncotarget*, vol. 7, pp. 20507–20519, 2016.

[22] Q. Xia, J. Xu, H. Chen et al., "Association between an elevated level of HMGBI and non-small-cell lung cancer: A meta-analysis and literature review," *OncoTargets and Therapy*, vol. 9, pp. 3917–3923, 2016.

[23] D. Musumeci, G. N. Roviello, and D. Montesarchio, "An overview on HMGB1 inhibitors as potential therapeutic agents in HMGB1-related pathologies," *Pharmacology & Therapeutics*, vol. 141, no. 3, pp. 347–357, 2014.

[24] L. Mollica, F. de Marchis, A. Spitaleri et al., "Glycyrrhizin binds to high-mobility group box 1 protein and inhibits its cytokine activities," *Chemistry & Biology*, vol. 14, no. 4, pp. 431–441, 2007.

[25] M. Ogiku, H. Kono, M. Hara, M. Tsuchiya, and H. Fujii, "Glycyrrhizin prevents liver injury by inhibition of high-mobility group box 1 production by kupffer cells after ischemia-reperfusion in rats," *The Journal of Pharmacology and Experimental Therapeutics*, vol. 339, no. 1, pp. 93–98, 2011.

[26] K. Xiang, L. Cheng, Z. Luo et al., "Glycyrrhizin suppresses the expressions of HMGB1 and relieves the severity of traumatic pancreatitis in rats," *PLoS ONE*, vol. 9, no. 12, Article ID e115982, 2014.

[27] R. Smolarczyk, T. Cichoń, S. Matuszczak et al., "The role of glycyrrhizin, an inhibitor of HMGB1 protein, in anticancer therapy," *Archivum Immunologiae et Therapia Experimentalis*, vol. 60, no. 5, pp. 391–399, 2012.

[28] J. F. Curtin, "HMGB1 mediates endogenous TLR2 activation and brain tumor regression," *PLOS Medicine*, vol. 6, p. e10, 2009.

[29] J. Elsam, "Histological and histochemical methods: theory and practice, 5th edition," *Biotechnic & Histochemistry*, vol. 91, no. 2, pp. 145-145, 2016.

[30] E. Sabattini, K. Bisgaard, S. Ascani et al., "The EnVision(TM)+ system: A new immunohistochemical method for diagnostics and research. Critical comparison with the APAAP, ChemMate(TM), CSA, LABC, and SABC techniques," *Journal of Clinical Pathology*, vol. 51, no. 7, pp. 506–511, 1998.

[31] Y. Zhang, J. Dong, P. He et al., "Genistein inhibit cytokines or growth factor-induced proliferation and transformation phenotype in fibroblast-like synoviocytes of rheumatoid arthritis," *Inflammation*, vol. 35, no. 1, pp. 377–387, 2012.

[32] J. S. Rawlings, K. M. Rosler, and D. A. Harrison, "The JAK/STAT signaling pathway," *Journal of Cell Science*, vol. 117, no. 8, pp. 1281–1283, 2004.

[33] C. Schindler, D. E. Levy, and T. Decker, "JAK-STAT signaling: from interferons to cytokines," *The Journal of Biological Chemistry*, vol. 282, no. 28, pp. 20059–20063, 2007.

[34] D. S. Aaronson and C. M. Horvath, "A road map for those who don't know JAK-STAT," *Science*, vol. 296, no. 5573, pp. 1653–1655, 2002.

[35] B. Lu, D. J. Antoine, K. Kwan et al. et al., "JAK/STAT1 signaling promotes HMGB1 hyperacetylation and nuclear translocation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, pp. 3068–3073, 2014.

[36] C. Ma, Y. Wang, L. Dong, M. Li, and W. Cai, "Anti-inflammatory effect of resveratrol through the suppression of NF-κB and JAK/STAT signaling pathways," *Acta Biochimica et Biophysica Sinica*, vol. 47, pp. 207–213, 2015.

[37] X. Wang, L. Xiang, H. Li et al., "The role of HMGB1 signaling pathway in the development and progression of hepatocellular carcinoma: a review," *International Journal of Molecular Sciences*, vol. 16, no. 9, pp. 22527–22540, 2015.

[38] P. Huebener, J.-P. Pradere, C. Hernandez et al., "The HMGB1/RAGE axis triggers neutrophil-mediated injury amplification following necrosis," *The Journal of Clinical Investigation*, vol. 125, no. 2, pp. 539–550, 2015.

Hindawi
BioMed Research International
Volume 2017, Article ID 5482750, 13 pages
https://doi.org/10.1155/2017/5482750

Research Article

An Improved Binary Differential Evolution Algorithm to Infer Tumor Phylogenetic Trees

Ying Liang, Bo Liao, and Wen Zhu

College of Information Science and Engineering, Hunan University, Changsha, China

Correspondence should be addressed to Bo Liao; dragonbw@163.com

Received 2 September 2017; Accepted 18 October 2017; Published 27 November 2017

Academic Editor: Tao Huang

Tumourigenesis is a mutation accumulation process, which is likely to start with a mutated founder cell. The evolutionary nature of tumor development makes phylogenetic models suitable for inferring tumor evolution through genetic variation data. Copy number variation (CNV) is the major genetic marker of the genome with more genes, disease loci, and functional elements involved. Fluorescence in situ hybridization (FISH) accurately measures multiple gene copy number of hundreds of single cells. We propose an improved binary differential evolution algorithm, BDEP, to infer tumor phylogenetic tree based on FISH platform. The topology analysis of tumor progression tree shows that the pathway of tumor subcell expansion varies greatly during different stages of tumor formation. And the classification experiment shows that tree-based features are better than data-based features in distinguishing tumor. The constructed phylogenetic trees have great performance in characterizing tumor development process, which outperforms other similar algorithms.

1. Introduction

Cancer is the most serious and dangerous disease to human health in the world. Over the past few decades, researchers have been working on the diagnosis and treatment of cancer. Owing to these great efforts, our understanding of cancer has been greatly improved, and early clinical diagnosis and reliable treatment are critical for cancer [1]. Cancer is the result of an imbalance in the cell cycle of the organism. Each cell of the organism contains a complete genome and has great spontaneity [1]. When the genome is no longer regulated by normal tissue and the spontaneity of cells is activated, then cancer develops. Tumor cells succumb to different evolutionary pressures and result in constant replication, growth, invasion, and metastasis [1].

In the early days, Nowell [2] proposed the "clonal evolution" theory that combines evolutionary biology with tumor biology. The model suggests a tumor is most likely to start with a mutated cell. Owing to the expansion of one or more cell subclones, tumor cells show high heterogeneity, which is an important characteristic of tumor development [3]. These tumor cells show significant differences even in the same tissue of the same individual. It has been shown that tumor heterogeneity is evolving along with tumor progression [3]. Tumor heterogeneity has been shown to have a significant impact on the diagnosis and treatment of cancer [3, 4].

Because of the evolutionary nature of tumor development, phylogenetic models were used to infer tumor evolution through genetic variation data [5]. Navin et al. [6] found that a single breast tumor may contain multiple cell subclones, and their chromosome copy numbers vary considerably via single-cell DNA copy number data on CGH platform. The development of next-generation sequencing allows people to infer SNVs and their allele frequencies in heterogeneous tumor cell populations. Because of the huge number of SNVs, inference of a complete tumor progression model to explain the observed data has encountered computational difficulties. Nik-Zainal et al. [7] reconstructs phylogenetic tree from inferred SNV frequencies based on two assumptions: (i) no mutation occurs twice in the course of cancer evolution and (ii) no mutation is ever lost. Strino et al. [8] proposed a linear algebra approach based on the two hypotheses to limit the number of possible trees, which can handle up to 25 SNVs. Detection of clones based on SNV frequency data is necessary for inferring phylogeny. Jiao et al. [9] proposes PhyloSub, a Bayesian nonparametric model,

to infer the phylogeny and genotype of the major subclonal lineages represented in the population of cancer cells. Miller et al. [10] proposed a variational Bayesian mixture model to identify the number and genetic composition of subclones by analyzing the variant allele frequencies. Hajirasouliha et al. [11] formulate the problem of constructing the subpopulations of tumor cells from the variant allele frequencies (VAFs) as binary tree partition and present an approximation algorithm to solve the max-BTP problem. El-Kebir et al. [12] formulate the problem of reconstructing the clonal evolution of a tumor using SNV as the VAF factorization problem and derives an integer linear programming solution to the VAF factorization problem. Popic et al. [13] propose LICHeE, a novel method to infer the phylogenetic tree of cancer progression from multiple somatic samples. Because of copy number alterations, loss of heterozygosity (LOH), and normal contamination, the allele frequencies of related SNV need to be corrected [14]. Copy number variation is segment loss or duplication of genome sequence ranging from kilo bases (Kb) to mega bases (Mb) in size, which covers 360 Mb and encompasses hundreds of genes, disease loci, and functional elements [15]. CNVs affect gene expressions in human cell-lines, which also play a major role in cancer [16]. Subramanian et al. [17] develop a novel pipeline for building trees of tumor evolution from the unmixed tumor copy number variations (CNVs) data. Oesper et al. [18] introduce ThetA, an algorithm to infer the most likely collection of genome and its proportions in a sample, and identify subclonal CNVs using high-throughput sequencing data. Ha et al. [19] also present a novel probabilistic model, TITAN, to infer CNA and LOH events while accounting for mixtures of cell populations, thereby estimating the proportion of cells harboring each event. Some tumor progression analysis tools combine VAFs of SNVs and population frequencies of structure variations to reconstruct subclonal composition and tumor evolution. PhyloWGS [20] uses copy number alterations to correct the VAFs of affected SNVs and greatly improves subclonal reconstruction compared to existing methods. As tumor is a heterogeneity system, Jiang et al. [21] propose Canopy to identify cell populations and infer phylogenies using both somatic copy number alterations and single-nucleotide alterations from one or more samples derived from a single patient. Li and Xie [22] propose a software package called PyLOH to deconvolve the mixture of normal and tumor cells using copy number alterations and LOH information. Yu et al. [23] introduce CloneCNA to address normal cell contamination, tumor aneuploidy, and intratumor heterogeneity issues and automatically detect clonal and subclonal somatic copy number alterations from heterogeneous tumor samples. El-Kebir et al. [24] develop SPRUCE to construct phylogenetic trees jointly from SNVs and CNAs, which overcomes complexities in simultaneous analysis of SNVs and CNAs.

The samples of the above studies are mixture of cancer cells and stromal cells; analyzing single cells is the most informative approach to assess the heterogeneity within a tumor [5]. Single-cell analysis is not only one more step towards more-sensitive measurements, but also a decisive jump to a more-fundamental understanding of biology [25]. Navin et al. [26] obtain robust high-resolution copy number profiles by sequencing a single cell and infer about the evolution and spread of cancer by examining multiple cells from the same cancer with the Euclidean metric. Traditionally used Euclidean or correlation distances for tree reconstruction from copy number profiles are ill-suited, owing to the dependent and nonidentical distribution of rearrangement events [5]. Fluorescence in situ hybridization (FISH) is a technique that can be used to count the copy number of DNA probes for specific genes or chromosomal regions in potentially hundreds of individual cells of a tumor. Pennington et al. [27] develop a new method combined with expectation maximization to infer unknown parameters for identifying common tumor progression pathways by taking advantage of information on tumor heterogeneity lost to prior microarray-based approaches on a set of fluorescent in situ hybridization (FISH) data. Chowdhury et al. [28–30] propose a software FISHtrees to build evolutionary trees of single tumors with FISH data. FISHtrees models gain or loss of genetic regions at the scale of single genes, whole chromosomes, or the entire genome, including variable rates for different gain and loss events in tumor evolution [30]. Later, Gertz et al. [31] present FISHtrees 3.0, which implements a ploidy-based tree building method based on mixed integer linear programming. The ploidy-based modeling in FISHtrees 3.0 includes a new formulation of the problem of merging trees for changes of a single gene into trees modeling changes in multiple genes and the ploidy [31]. Here, we propose an improved binary differential evolution algorithm to infer phylogenetic trees (BDEP) using CNV data of cervical cancer and breast cancer. The cervical cancer dataset contains the copy number profiles of four genes, and breast cancer dataset is up to eight genes. Liu et al. [32] show that, on average, each cancer can be explained with around six different marker sets. Tumor phylogenetic tree inference can be treated as minimum Steiner tree problem in directed graph, which is a NP-hard problem. BDEP uses differential individual to search for the best approximate solutions, with the help of individual's difference information and neighborhood optimal information to update. BDEP overcomes the weakness that differential evolution algorithm can only be used in continuous search space with advantages of fast convergence and strong robustness.

2. Methods

2.1. Problem Definition. One copy number variation usually affects the copy number of two or more closely related genes [15]. The genes may change their copy number alone or together with their neighbors located in one copy number variation region, which results in computational difficulties of evolution distance between gene copy number profiles (Figure 1). Shamir et al. propose an algorithm that calculates evolution events in linear time and linear space by backtracking the dynamic programming vector [33]. We adopt the idea proposed by Shamir to calculate the minimum variation events between two copy number profiles. Profiles (u, v) present the evolution distance from the source profile u to the target profile v. As mentioned by Shamir et al. [33],

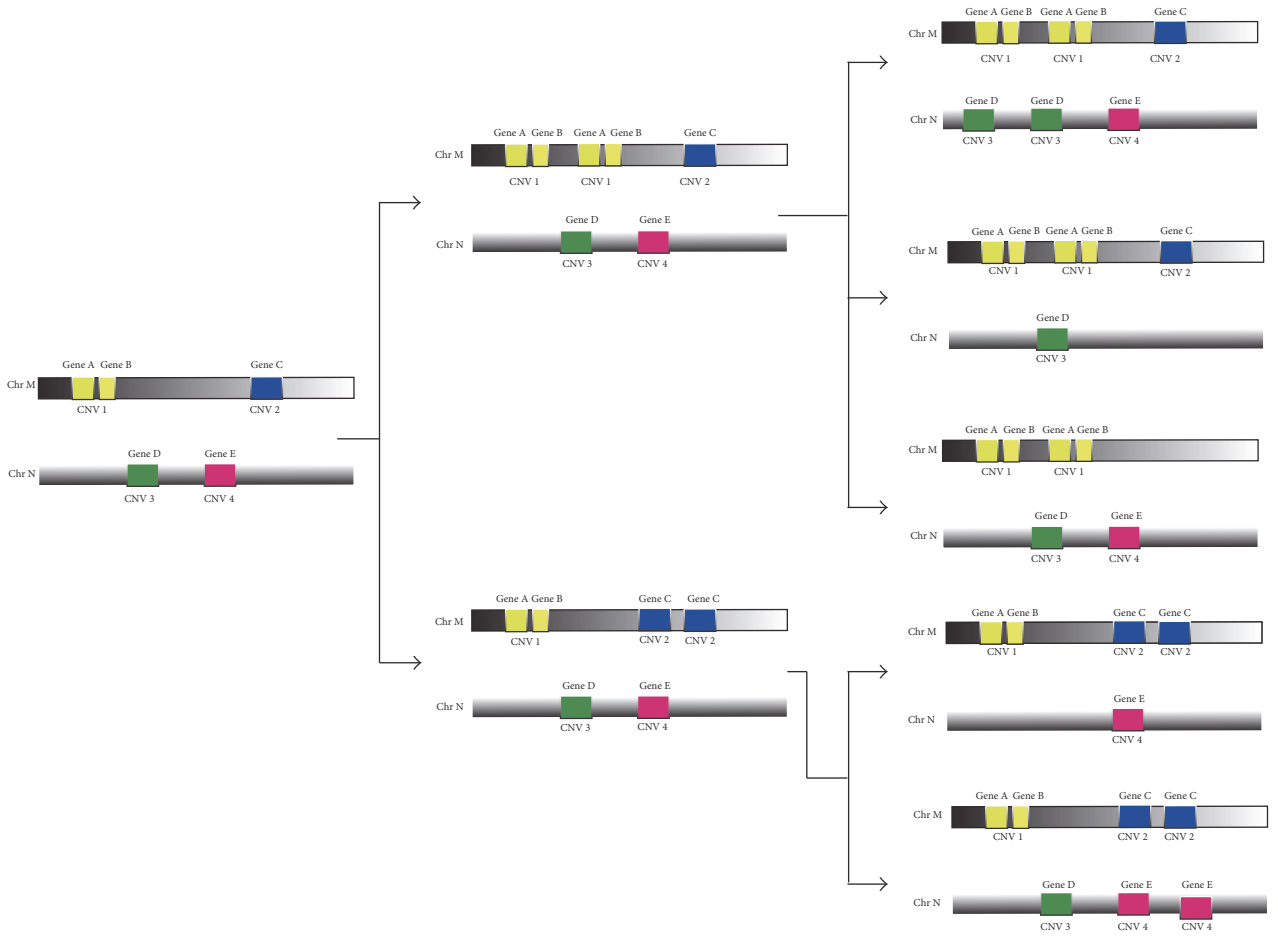


Figure 1: The association between CNVs and genes.

if the source profile contains the gene with copy number 0 but the target profile with the gene copy number > 0, the transformation from u to v is unreachable. On the contrary, if the gene has copy number > 0 in the source profile but with the copy number 0 in the target profile, the profiles (u, v) can be inferred. The distance matrix between copy number profiles is asymmetric, which corresponds to directed edges between copy number profiles.

Cells are continuously growing, proliferating, and dying during the tumor progress; the dying cells disappeared but once played an important role in tumourigenesis. Construct a tree to describe evolutionary relationship of observed cells and dying cells can be regarded as Steiner tree problem; the dying cells in Steiner tree are Steiner node. The Steiner tree problem is a classical combinatorial optimization problem, which has important applications in the fields of computer network layout, circuit design, and biological network analysis. In the paper, the tumor phylogenetic tree is a Steiner minimum tree problem in graph, which is proposed by Hakimi [34] and Hwang et al. [35]. The problem can be described as follows: Given a directed connected graph $G = (V, E)$ with observed nodes and all possible Steiner nodes, V, and edges, E, each node presents a copy number profile and each edge presents the evolution direction between nodes. The weight of each edge presents the evolution distance between copy number profiles. There is a subset $P \subseteq V$; each element presents the observed copy number profile of cell. The Steiner tree problem is to find a subtree T of directed connected graph G, which contains all nodes in P with minimal weight sum. The subtree T is the Steiner tree of subset P; the node that exists in T but not in P is the Steiner node. When $P = V$, the Steiner tree problem is minimum arborescence problem, which can be worked out in polynomial time [36]. Otherwise, the Steiner tree problem has no polynomial time solution, which is a NP-hard problem [37]. When the input scale becomes large, it is impossible to find the exact optimal solution in polynomial time. Therefore, a good approximation algorithm will provide a compromise solution for the NP-hard problem.

2.2. The Improved Binary Differential Evolution Model. The differential (DE) evolution algorithm does not depend on the characteristics information of problem, with the help of difference information among individuals to disturb the formation of individual and then to search the entire population space. Greedy competition mechanism is employed to seek the optimal solution of the problem. DE algorithm is a population-based stochastic direct search method, which is based on real number coding [38]. The differential evolution algorithm has the advantages of fast convergence, simple

operation, easy programming, and strong robustness, which have been widely used in various fields [39–42]. The DE algorithm contains three basic operations: mutation, crossover, and selection. The initial population is randomly generated and covers the entire search space.

Initial Population. Suppose $X_{i,G} = \{x_{i,G}^1, \ldots, x_{i,G}^n\}$ is the ith individual of generation Gth; n is the dimension of individual; $i = 1, 2, \ldots, M$ is the population scale; $G = 1, 2, \ldots, G_{\max}$ is the maximum evolution generation. The initial population of DE is generated by

$$x_{i,0}^j = \text{rand}_j(0,1)\left(x_U^j - x_L^j\right) + x_L^j, \tag{1}$$

where x_U^j and x_L^j represent the upper and lower bounds of the jth dimension, respectively, and $\text{rand}_j(0,1)$ represents a random number within the range $[0, 1]$.

Mutation Operation. Randomly select two different individuals $X_{p_1,G}$, $X_{p_2,G}$ to produce the mutant individual $V_{i,G}$ corresponding to individual $X_{i,G}$ as

$$v_{i,G}^j = x_{i,G}^j + \lambda\left(x_{p_1,G}^j - x_{p_2,G}^j\right), \tag{2}$$

where $x_{p_1,G}^j - x_{p_2,G}^j$ is difference vector and scaling factor λ is a positive control parameter of difference vector.

Crossover Operation. Crossover operation aims at increasing population diversity. The crossover strategy exchanges mutant and old individual's information to generate trial individual $U_{i,G}$. The crossover operation is defined as

$$u_{i,G}^j = \begin{cases} v_{i,G}^j & \text{rand}_j[0,1) \le \text{CR or } j = \text{rand}(i) \\ x_{i,G}^j & \text{otherwise.} \end{cases} \tag{3}$$

The crossover strategy ensures that $U_{i,G}$ has at least one element from $V_{i,G}$. The crossover rate CR can be adjusted by user within the range $[0, 1]$.

Selection Operation. Trial individual $U_{i,G}$ will become a member of the next-generation population, if the fitness function values of $U_{i,G}$ are superior to $X_{i,G}$. Otherwise, the individual $X_{i,G}$ will remain in the next-generation population. The selection operation is defined as

$$X_{i,G+1} = \begin{cases} U_{i,G}, & \text{fitness}\left(U_{i,G}\right) \le \text{fitness}\left(X_{i,G}\right) \\ X_{i,G}, & \text{otherwise.} \end{cases} \tag{4}$$

Perform the above three operations repeatedly until the stopping criterion is satisfied.

2.2.1. Binary Differential Evolution Algorithm. Conventional DE algorithm focuses on the problem of continuous search space, which cannot solve the discrete problem. Also the DE algorithm does not take into account the global or neighborhood optimal individual information. In this paper, we propose a novel binary differential evolution algorithm (BDEP) to solve the Steiner tree problem and further construct tumor phylogenetic tree. In BDEP, trial individual absorbs neighborhood optimal individual information to update at crossover phase. BDEP is different from conventional DE algorithm at initial population operation, mutation operation, and crossover operation. The algorithm flow chart of BDEP is in Algorithm 1.

Candidate Steiner Node Generation. The Steiner tree problem in graph is to find a minimum arborescence which at least contains all nodes in subset P. The set of nodes V in graph G includes the nodes in P and all possible Steiner nodes. Before applying Chu-Liu's algorithm to find the minimum arborescence, it is prerequisite to compute all possible Steiner points. The candidate Steiner node is generated according to the gene copy number profile in subset P. Under maximum parsimony criterion, the evolutionary distance from gene copy number profile to the candidate Steiner node is 1. As a result, the set of nodes V consists of candidate Steiner nodes and subset P, which corresponds to a complete directed graph G.

Individual Encoding. The individual i of binary differential evolution is encoded as a binary string $X_i = (x_i^1, x_i^2, \ldots, x_i^n)$, where x_i^j is a binary variable corresponding to the jth candidate Steiner node and n is the number of candidate Steiner nodes. When $x_i^j = 1$, the ith individual has the jth candidate Steiner node. With the gene copy number profile in set P, each individual represents a phylogenetic tree; the fitness function is the distance sum of the phylogenetic tree. The objective of BDEP is to find a minimum arborescence representing tumor phylogenetic tree.

Initial Population. The population initialization of BDEP is as follows:

$$x_{i,0}^j = \begin{cases} 1 & \text{rand}_j(0,1) < 0.05 \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

The meaning of i, j, and $\text{rand}_j(0,1)$ is the same as that of conventional DE algorithm.

Mutation Operation. For each individual $X_{i,G}$, randomly select two different individuals $X_{p_1,G}$, $X_{p_2,G}$ to produce the mutant individual $V_{i,G}$ as follows:

$$v_{i,G}^j = \begin{cases} x_{p_1,G}^j \mid x_{p_2,G}^j & x_{p_1,G}^j = x_{p_2,G}^j \\ x_{i,G}^j & \text{otherwise.} \end{cases} \tag{6}$$

For the jth candidate Steiner node, if individuals $X_{p_1,G}$, $X_{p_2,G}$ have the same choice, the mutant individual yields $x_{p_1,G}^j$ or $x_{p_2,G}^j$; otherwise it directly derives from $X_{i,G}$.

Crossover Operation. Social learning is an important way to improve population diversity and self-adaptability. The individual would influence its neighbors: BDEP uses local neighborhood as social learning areas. BDEP adopts the ring

Require: The copy number profiles (object nodes set P).
The max generation $G_{\max}$.
The number of individuals (population scale) M.
Ensure: The tumor Steiner tree with the shortest length.
(1) Generate candidate Steiner node according to the copy number profiles, construct a complete directed graph *Graph*.
(2) Set the generation number $G \leftarrow 0$, initialize a population of M individuals $P_G = \{X_{1,G}, \ldots, X_{M,G}\}$ with $X_{i,G} = \{x^1_{i,G}, \ldots, x^n_{i,G}\}$ where $x^n_{i,G} \in \{0, 1\}$ is a binary variable.
(3) **while** stopping criterion is not satisfied **do**
(4) **Mutation step**
(5) **for** $i \leftarrow 1$ to M **do**
(6) Generate a mutant individual $V_{i,G} = \{v^1_{i,G}, \ldots, v^n_{i,G}\}$ from the target individual $X_{i,G}$ and two different individuals $X_{p1,G}, X_{p2,G}$.
(7) **for** $j \leftarrow 1$ to n **do**
(8)
$$v^j_{i,G} = \begin{cases} x^j_{p1,G} \text{ or } x^j_{p2,G} & x^j_{p1,G} = x^j_{p2,G} \\ x^j_{i,G} & \text{otherwise} \end{cases}$$
(9) **end for**
(10) **end for**
(11) **Crossover step**
(12) **for** $i \leftarrow 1$ to M **do**
(13) Search the r-neighborhood of individual $V_{i,G}$, the best neighbor of $V_{i,G}$ is $V_{n\text{best},G} = \min_{r\text{-neighborhood}} \text{fitness}$
(14) Update trial individual $V_{i,G}$ to $U_{i,G}$
(15) $\text{rand}(i) = \lfloor \text{rand}[0, 1) * n \rfloor$
(16) **for** $j \leftarrow 1$ to n **do**
(17)
$$u^j_{i,G} = \begin{cases} v^j_{n\text{best},G} & \text{rand}[0.1) \le \text{CR or } j = \text{rand}(i) \\ v^j_{i,G} & \text{otherwise} \end{cases}$$
(18) **end for**
(19) **end for**
(20) **Selection step**
(21) **for** $i \leftarrow 1$ to M **do**
(22) Evaluate the trial individual $U_{i,G}$
(23) **if** $\text{fitness}(U_{i,G}) \le \text{fitness}(X_{i,G})$ **then**
(24) $X_{i,G+1} = U_{i,G}$, $\text{fitness}(X_{i,G+1}) = \text{fitness}(U_{i,G})$
(25) **end if**
(26) **end for**
(27) **Update the generation count** $G \leftarrow G + 1$
(28) **end while**
(29) **return** optimal tumor Steiner tree T

Algorithm 1: An improved binary differential evolution algorithm to infer tumor phylogenetic trees (BDEP).

topology of population with radius r to define local neighborhoods. The r-neighborhood of individual i is represented as $\{R_j \mid |i - j| \le r,\ j = 0, 1, 2, \ldots, M - 1\}$. The individual $V_{n\text{best},G}$ represents the best neighbors with minimum fitness value in the r-neighborhood of mutant individual $V_{i,G}$. The cross operation is according to

$$u^j_{i,G} = \begin{cases} v^j_{n\text{best},G} & \text{rand}_j [0, 1) \le \text{CR or } j = \text{rand}(i) \\ v^j_{i,G} & \text{otherwise.} \end{cases} \quad (7)$$

The crossover strategy exchanges mutant individual and its best neighbor's information to generate trial individual. The crossover rate CR can be adjusted by user within the range $[0, 1]$. The crossover strategy ensures that $U_{i,G}$ has at least one element from the best neighbor. The neighborhood radius r depends on population scale and the complexity of problem.

Selection Operation. The selection strategy is similar to conventional DE algorithm; whether the trial individual $U_{i,G}$ could become a member of the next-generation population depends on fitness function values. If the new individual $U_{i,G}$ is superior to old one $X_{i,G}$, $U_{i,G}$ would replace $X_{i,G}$. Otherwise, the individual $X_{i,G}$ will remain in the next-generation population.

Repeatedly perform the above three operations until one of the two criteria is satisfied: (i) evolutional iterations reach the maximal generation; (ii) the optimal fitness value is less than the distance sum of minimum arborescence of subset P and stays unchanged in ten consecutive iterations.

3. Results and Discussion

In this section, we apply BDEP to the gene copy number profiles of real tumor and infer the tumor phylogeny of

Table 1: The P value of χ tests between DCIS and IDC.

Sample ID	P value of branches	P value of levels	P value of edges
Patient 1	$4.89E-56$	$8.40E-03$	$5.85E-01$
Patient 2	$4.49E-34$	$5.61E-20$	$9.25E-01$
Patient 3	$1.82E-03$	$1.38E-02$	$8.91E-01$
Patient 4	$5.53E-41$	$1.86E-06$	$2.24E-02$
Patient 5	$2.24E-18$	$4.28E-03$	$5.81E-01$
Patient 6	$4.87E-20$	$5.22E-02$	$3.14E-03$
Patient 7	$6.11E-02$	$1.06E-05$	$1.40E-01$
Patient 8	$2.79E-61$	$1.45E-20$	$2.88E-01$
Patient 9	$1.09E-36$	$1.50E-18$	$7.94E-01$
Patient 10	$6.05E-58$	$1.38E-11$	$9.61E-01$
Patient 11	$1.30E-04$	$5.96E-16$	$8.29E-02$
Patient 12	$7.85E-02$	$7.40E-06$	$4.59E-01$
Patient 13	$2.43E-14$	$4.01E-05$	$9.32E-01$

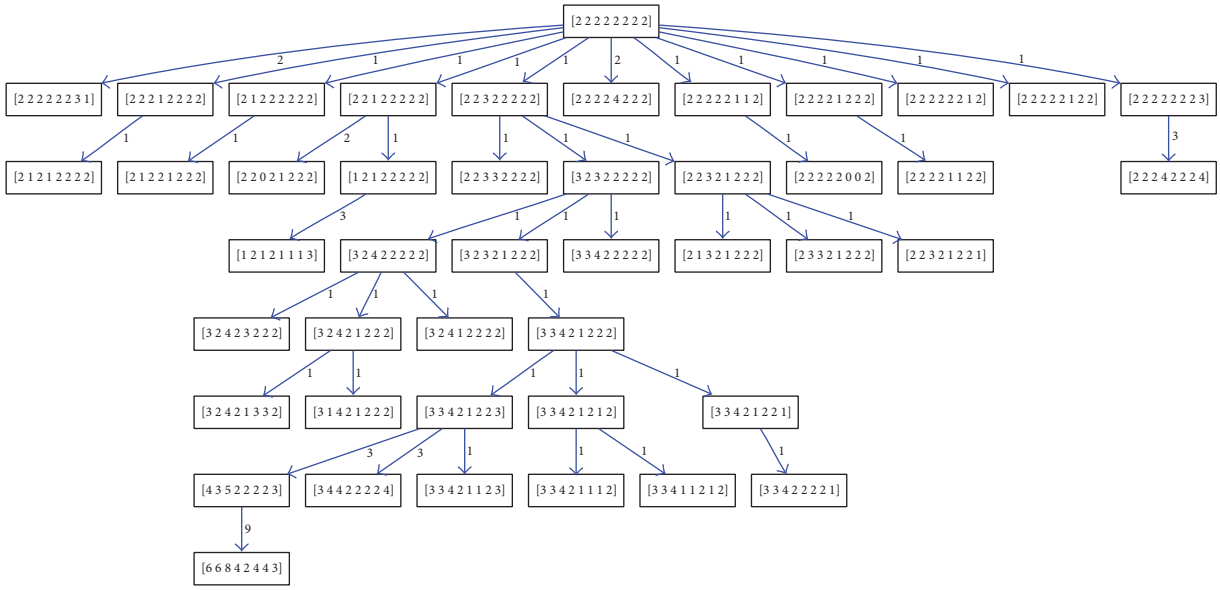
all samples. We study the differences between tumors by statistically analyzing topological features of phylogenetic tree in the following three aspects: branch, level, and edge. And classification experiments are performed to evaluate the merits of these features. The algorithm parameters are set as follows: the max generation $G_{\max}$ is 100; crossover rate (CR) is 0.7 by default; and population size depends on the complexity of the problem ranging from 300 to 500.

3.1. Datasets. Two FISH datasets, cervical cancer and breast cancer, respectively, from Wangsa et al. [43] and Heselmeyer-Haddad et al. [44], are published to visualize copy number changes in tumors based on single-cell analyses. The cervical cancer dataset comprises four probes targeting the genes LAMP3, PROX1, PRKAA1, and CCND1, in pretreatment cervical biopsies from 16 lymph node positive samples and 15 lymph node negative controls from women with stage IB and IIA cervical cancer [43]. The lymph node positive samples contain primary tumors and associated lymph node metastases. The four target genes come from different chromosomes: LAMP3 is a gene located on chromosome 3q26, PROX1 is located on chromosome 1q41, PRKAA1 is located on chromosome 5p19, and CCND1 is located on chromosome 11q13; and altered expression of this gene has been observed in many cancers [43]. The cell number of cervical cancer among 47 cases ranges from 212 to 250 (average cell number is 243), which is not significantly different among primary cancer with positive lymph node, lymph node metastases cases, and lymph node negative controls. But the number of cell gene profiles among them is strikingly different; each gene copy number profile is a tree node in phylogenetic model. The gene profile number of primary cases with positive lymph node ranges from 63 to 187, average being 111. The profile number of lymph node metastases cases ranges from 34 to 115, average being 70. The profile number of lymph node negative controls ranges from 58 to 157, average being 97.

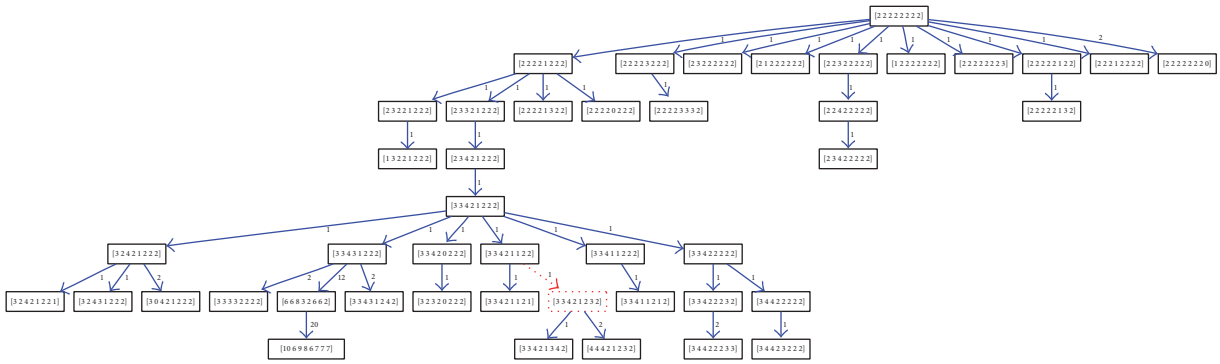
The breast cancer dataset comprises 13 cases of synchronous ductal carcinoma in situ (DCIS) and invasive ductal carcinoma (IDC), which contains eight probes targeting five oncogenes, COX2, MYC, HER2, CCND1, and ZNF217, and three tumor suppressor genes, DBC2, CDH1, and TP53 [44]. The DCIS is considered a precursor lesion for invasive breast cancer, which has a lower degree of chromosomal instability than the IDC [44]. COX2 is located on 1q31.1 and is upregulated in human breast cancer; DBC2 and MYC both are located on chromosome 8; MYC is also upregulated gene in many types of cancers; CDH1 is located on 16q22.1, HER2 and TP53 both are located on chromosome 17, and ZNF217 is located on 20q13.2, which is a strong candidate oncogene for breast and other cancers [44]. The cell number of breast cancer among 26 cases ranges from 76 to 220, average cell number being 142. The cell number and profile number between DCIS and IDC cases are not significantly different. The profile number of DCIS cases ranges from 28 to 143, average being 73. The profile number of IDC cases ranges from 44 to 119, average being 85.

In FISH datasets, gene copy number profiles of each cell are expressed in matrix form, where each row represents a cell case and each column represents a gene probe. The corresponding gene copy number of each cell is a nonnegative integer. The profile with gene copy number of 2 is considered as the root node of tumor evolutionary tree. The datasets can be downloaded at ftp://ftp.ncbi.nlm.nih.gov/pub/FISHtrees/.

3.2. Results on Breast Cancer Datasets. We apply BDEP algorithms to the gene copy number profiles of breast cancer and comparatively analyze the tree topology between paired DCIS and IDC samples. We first analyze the branch features of phylogenetic tree at different stages. The branch is defined as subtree derived from the ith child of the root node. The DBC2 and MYC gene are on chromosome 8, and TP53 and HER2 gene are on chromosome 17. The copy number of genes lying on the same chromosome is easily affected by CNV simultaneously, phylogenetic trees have at most twenty branches, and we use Chi-square test to compare the distribution characteristics of cell numbers of each branch. The P values of Chi-square test from 13 paired samples are listed in Table 1. The P value of Chi-square test less than 0.01 is considered significant. For patients 7 and 12, the branch structures of phylogenetic tree are similar. But the branch



(a) The phylogenetic tree of ductal carcinoma in situ



(b) The phylogenetic tree of invasive ductal carcinoma

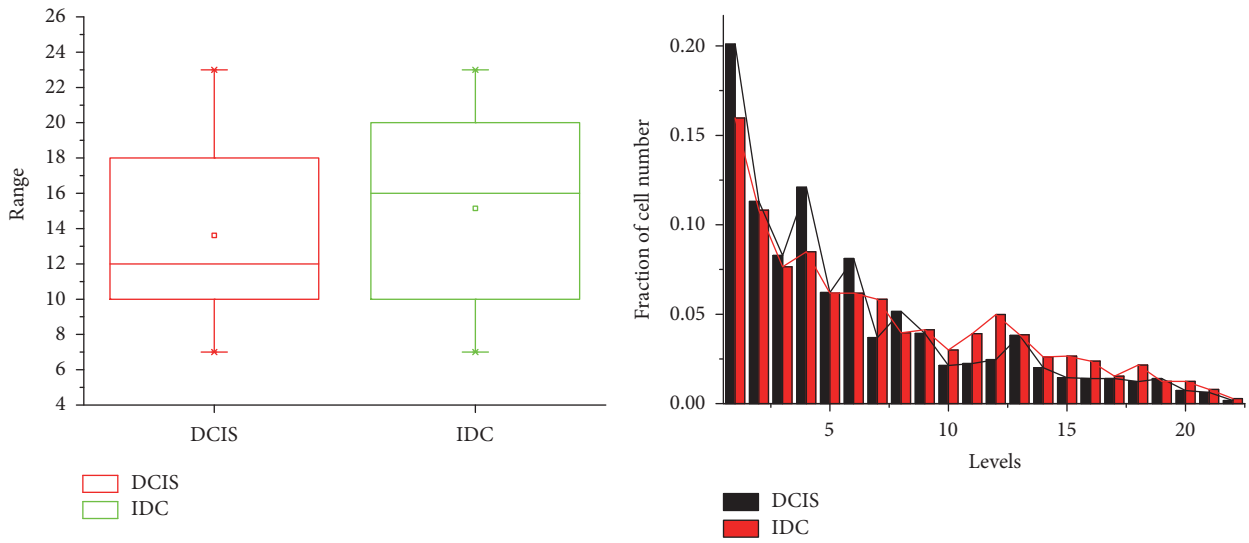
Figure 2: The comparison of BC phylogenetic trees.

structures of the remaining 11 paired samples are significantly different, which means that, under different selection pressures, the pathways of tumor subcellular amplification also change. As shown in Figure 2, which is an example of tumor phylogenetic tree from patient 5, Figures 2(a) and 2(b) are, respectively, from DCIS and IDC samples. The node in red is Steiner node and the weight is evolution distance between two nodes. The DCIS phylogenetic tree is more balanced, with more cells concentrated in the first four levels.

The cells number of phylogenetic tree across levels between DCIS and IDC tumor shows a noticeable difference. The P value of Chi-square test across the first twenty-two levels is listed in Table 1; the root node is on level zero. For the 13 paired samples, there are 11 cases with statistical significance. The hierarchical topology of primary and metastasis trees is similar in patients 3 and 6. We also analyze the depth characteristics of trees and corresponding fraction of cell number at each level. From Figure 3(a), the depth of DCIS tree is not distinctly different from IDC. The cell number distribution across different levels is illustrated in Figure 3(b). For the first six levels, the cell distribution of DCIS is more concentrated with a greater proportion compared with IDC. The cells gather in the first six levels up to 66% in DCIS and 55% in IDC. The number of cells decreases with the increment of tree levels, especially for DCIS. We also compare the edge features of phylogenetic trees; each edge is the corresponding gene gain or loss in the tree topological structure. The P value of edge statistics is not significantly different between DCIS and IDC except for patient 6, which is listed in Table 1.

3.3. Results on Cervical Cancer Datasets

3.3.1. Statistical Analysis of Tree Feature. BDEP is applied to comparatively analyze the tree topology between paired primary tumor and metastasis samples. The four genes of cervical cancer are on different chromosomes, phylogenetic trees have at most eight branches, and we use Chi-square test to compare the distribution characteristics of cell numbers of each branch. The Chi-square test of branch structure from 16 paired samples shows significant differences, which is listed in Table 2. The tree topology structure of primary and metastasis tumor is quite different. As shown in Figure 4, which is an



(a) The level count comparison of DCIS and IDC phylogenetic tree

(b) The cell number comparison of DCIS and IDC phylogenetic tree

Figure 3: The level characteristics of BC phylogenetic tree.

example of tumor phylogenetic tree from patient 3, Figures 4(a) and 4(b) are, respectively, from primary and metastasis samples. The node in red is Steiner node and the weight is evolution distance between two nodes. The metastasis sample has less copy number profiles, and the corresponding tree has fewer levels but with more balanced and broader topological structure compared with primary one.

In order to find the most decisive gene to distinguish primary and metastasis samples, we analyze the significance of individual gene. For each gene, we compare the cell numbers of branches with gene loss and gain. From Table 2, it is obvious that gene LAMP3 is the most informative gene; there are seven cases showing significant difference (patients 5, 6, 7, 12, 13, 14, and 16), which is consistent with the findings of Kanao et al. [45] and Mine et al. [46]. The overexpression of LAMP3 is associated with an enhanced metastatic potential and may be a prognostic factor for cervical cancer [45]. The gene PRKAA1 is the least with only two significant cases (patients 3 and 11).

For the hierarchical structure of trees, the P value of Chi-square test across the first twelve levels is listed in Table 3. Among the 16 paired samples, there are 14 cases with statistical significance. The hierarchical topology of primary and metastasis trees is distinguishable except for patients 1 and 9. The depth characteristics of trees and corresponding fraction of cell number at each level are illustrated in Figure 5. Whether or not lymph node later metastasized, the level structure of primary tumor is not distinctly different, but much deeper than the metastasized one. The cell distribution of metastasis sample is more concentrated and most of them gather in the first six levels compared with primary stage tumor. The number of cells decreases with the increment of tree levels, especially for metastasis tumor. The cells gather in the first six levels up to 85% in metastasis tumor and 70% in primary tumor. The cells in primary tumor are more evenly distributed and extending to more levels. For the edge feature of phylogenetic tree, all the 16 paired samples show no significant difference, which is similar to breast cancer samples.

For the edge feature of phylogenetic tree, all the 16 paired samples show no significant difference, which is similar to breast cancer samples.

3.3.2. The Classification Evaluation on Tree Features. The performance to predict the state of the tumor according to topological features of trees is crucial, which provides diagnostic guidance for accurate medical treatment. We evaluate the tree features through classification experiments and compare them with the features directly from data. We use the support vector machines (SVM) as classifier, which is implemented in an open source machine learning Scikit-learn module for Python [47]. We perform three classification experiments on CC dataset and the average accuracy of 100 tests is considered as experimental result. The three classification experiments are as follows:

(1) Distinguishing primary from its corresponding metastatic samples, which is a 16 versus 16 samples' classification

(2) Distinguishing nonmetastasis primary from primary samples, which is a 15 versus 16 samples' classification

(3) Distinguishing primary and nonmetastasis primary samples from metastatic samples, which is a 16 versus 15 versus 16 samples' classification.

The dataset is divided into four parts: three of them are training sets and the remaining one is test set. The extracted features from tree topology are branch, level, and edge. There are two features derived from data: (i) maximum copy number of each gene; (ii) average copy number of each gene. BDEP also compares with the published FISHtrees algorithm [30], which is a state-of-the-art algorithm for

Table 2: The P value of branches χ tests between primary and metastasis samples of cervical cancer.

Sample ID	P value	P value of LAMP3	P value of PROX1	P value of PRKAA1	P value of CCND1
Patient 1	$2.56E-15$	$7.86E-01$	$3.01E-06$	$2.16E-01$	$4.97E-01$
Patient 2	$6.87E-18$	$4.05E-02$	$7.49E-03$	$9.56E-01$	$6.32E-01$
Patient 3	$1.23E-48$	$8.71E-01$	$2.90E-01$	$2.22E-03$	$3.80E-48$
Patient 4	$1.00E-48$	$3.74E-01$	$1.55E-10$	$5.24E-02$	$1.41E-01$
Patient 5	$1.39E-17$	$4.65E-05$	$6.50E-02$	$5.00E-01$	$8.74E-01$
Patient 6	$1.20E-18$	$3.20E-09$	$6.01E-02$	$3.51E-02$	$4.48E-03$
Patient 7	$3.64E-28$	$1.96E-06$	$5.76E-01$	$9.55E-01$	$5.09E-02$
Patient 8	$8.17E-72$	$5.47E-01$	$1.99E-20$	$1.11E-01$	$3.45E-03$
Patient 9	$1.52E-30$	$6.03E-02$	$7.52E-02$	$8.10E-01$	$9.01E-01$
Patient 10	$8.15E-10$	$4.22E-02$	$6.22E-01$	$1.44E-01$	$9.26E-06$
Patient 11	$1.21E-31$	$5.65E-01$	$6.07E-01$	$1.84E-12$	$5.63E-05$
Patient 12	$6.98E-55$	$1.15E-26$	$1.41E-06$	$5.67E-01$	$7.89E-01$
Patient 13	$4.71E-73$	$6.11E-35$	$9.56E-01$	$1.89E-02$	$1.39E-03$
Patient 14	$2.70E-18$	$2.29E-06$	$1.48E-02$	$5.17E-02$	$1.20E-02$
Patient 15	$7.77E-22$	$6.39E-01$	$1.72E-03$	$2.36E-02$	$3.81E-01$
Patient 16	$1.19E-27$	$3.06E-03$	$8.23E-01$	$7.50E-01$	$3.53E-01$

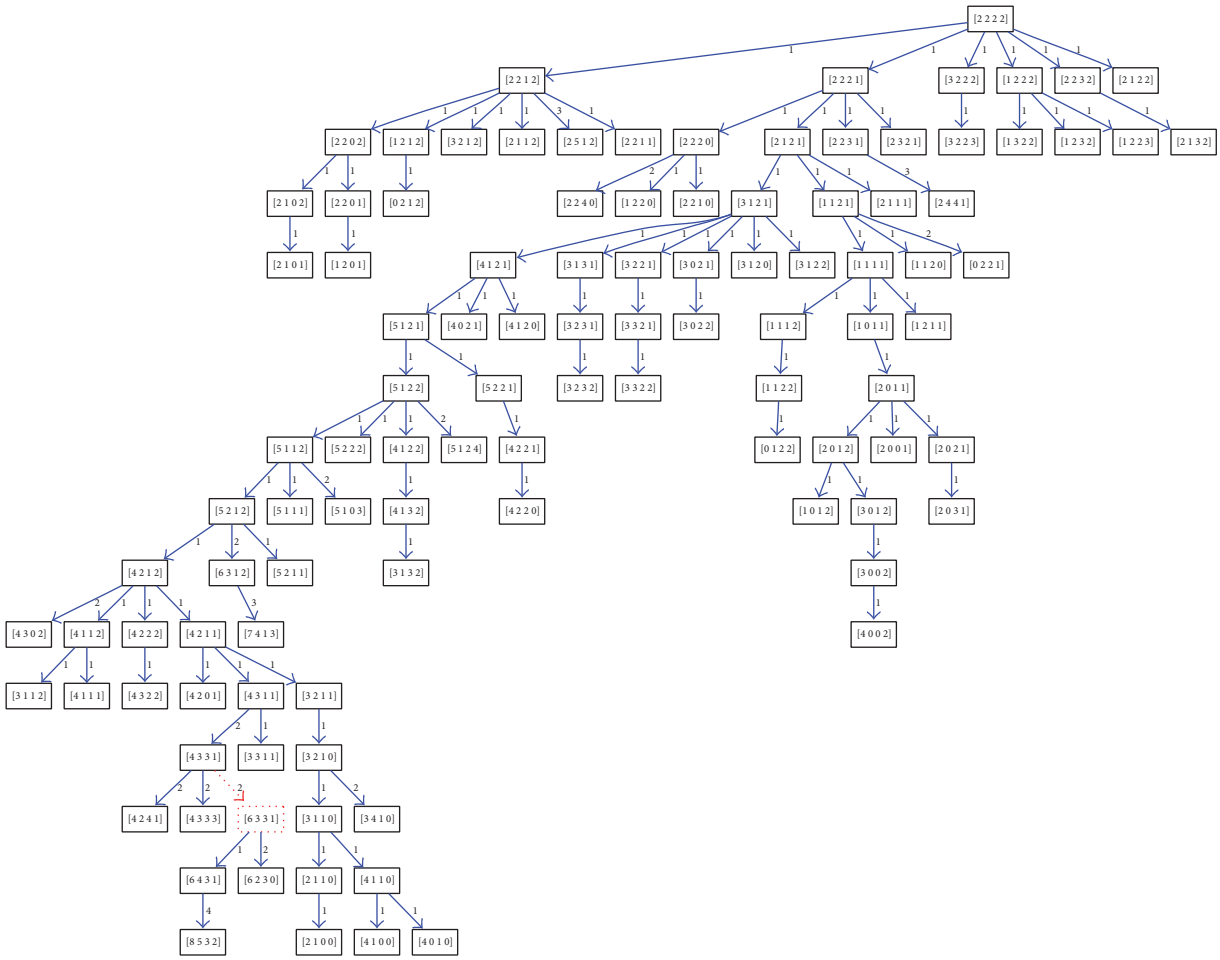
Table 3: The P value of levels and edges χ tests between primary and metastasis samples of cervical cancer.

Sample ID	P value of levels	P value of edges
Patient 1	$2.16E-02$	$9.35E-01$
Patient 2	$9.81E-09$	$6.48E-01$
Patient 3	$3.66E-17$	$8.04E-01$
Patient 4	$1.43E-05$	$9.06E-01$
Patient 5	$2.79E-07$	$3.34E-01$
Patient 6	$6.19E-09$	$6.82E-01$
Patient 7	$3.46E-04$	$9.64E-01$
Patient 8	$1.22E-07$	$7.97E-01$
Patient 9	$1.30E-02$	$9.25E-01$
Patient 10	$2.17E-09$	$8.28E-01$
Patient 11	$3.84E-10$	$4.98E-01$
Patient 12	$1.92E-15$	$2.49E-01$
Patient 13	$6.76E-17$	$2.87E-01$
Patient 14	$2.34E-06$	$6.75E-01$
Patient 15	$7.85E-03$	$6.48E-01$
Patient 16	$1.02E-16$	$9.90E-01$

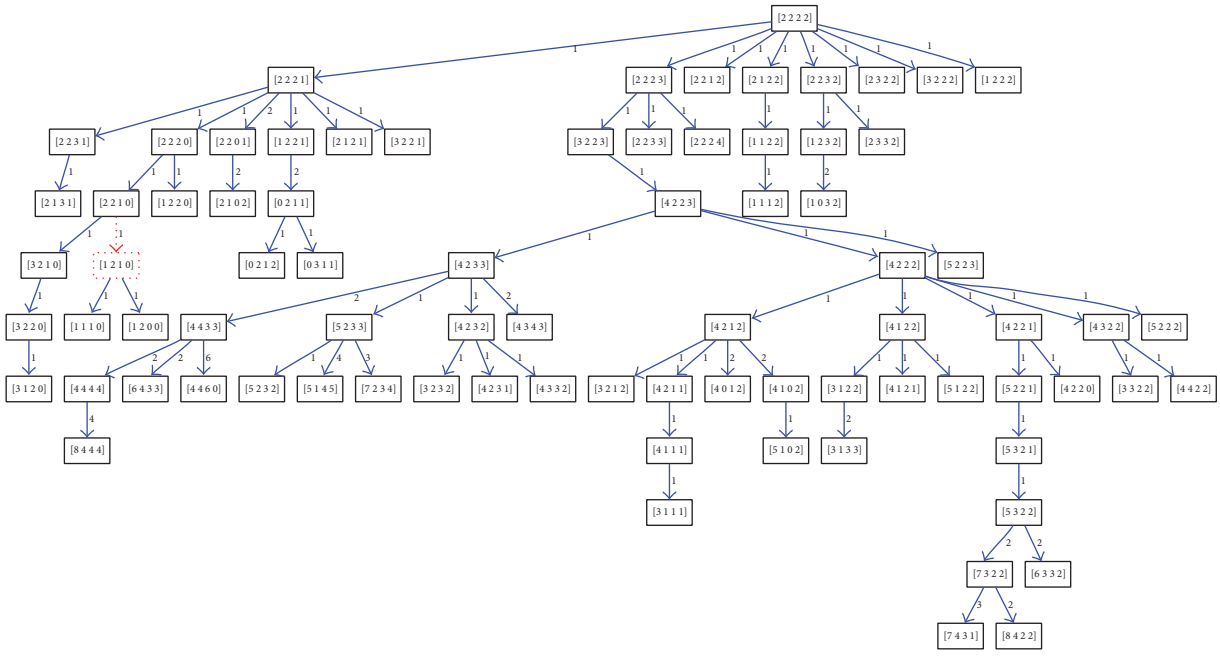
phylogenetic tree based on FISH platform; the result is shown in Figure 6. The experiment distinguishing primary from its corresponding metastatic samples works best, followed by the classification between primary samples. The effect of distinguishing primary, nonmetastasis primary, and metastatic samples is poor for all features. Among all the features, the level feature achieves the highest accuracy, which shows that the degree of cell differentiation varies widely for tumors of different states. The data-based average feature shows in general the worst performance. Also interestingly, the Chi-square tests of branch structure are significant for all 16 paired samples, but classification effect is not as good as expected, even worse than edge feature. FISHtrees works better than BDEP for branch structure feature, but not for edge and level features. Overall, the classification accuracy of tree-based feature is better than data-based feature.

4. Conclusion

In this paper, we propose a binary differential evolution algorithm (BDEP) to construct tumor phylogenetic tree via CNV data on FISH platform. Tumor phylogenetic tree inference can be treated as minimum Steiner tree problem in directed graph, which cannot be solved in polynomial time unless no Steiner node exists. The binary differential evolution is a heuristic algorithm with advantages of fast convergence and strong robustness, which provides good approximate solutions with reduced running time. Experimental results on real datasets show that the branch and hierarchical structures

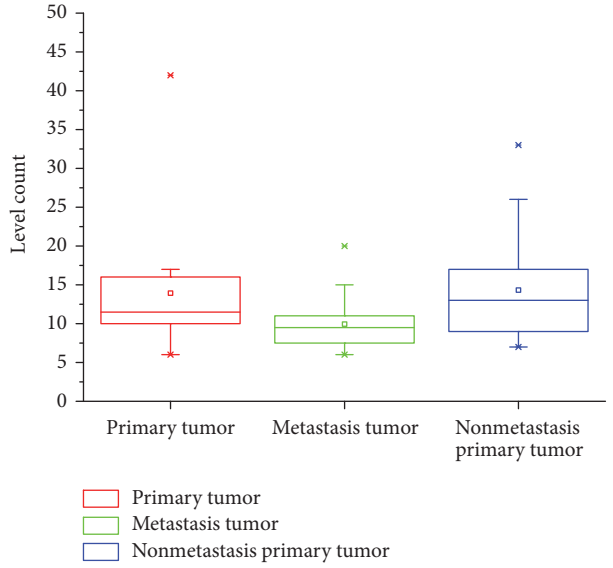


(a) The phylogenetic tree of primary cervical cancer

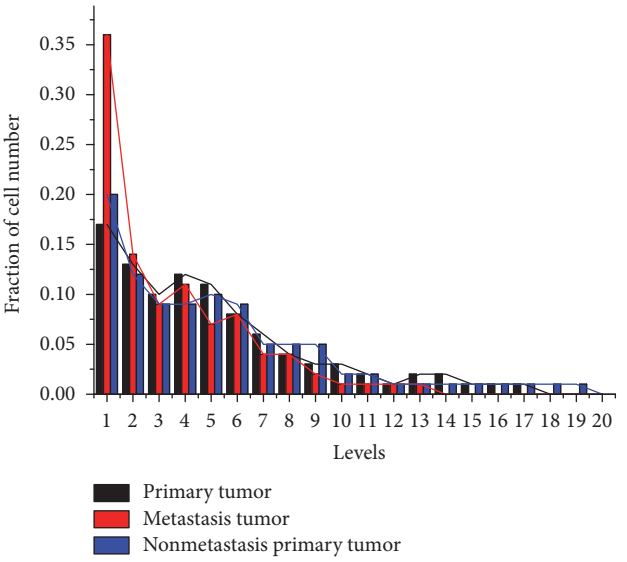


(b) The phylogenetic tree of lymph node metastasis cervical cancer

Figure 4: The comparison of CC phylogenetic trees.



(a) The level count comparison of primary, metastasis, and nonmetastasis primary tumor



(b) The cell number comparison of primary, metastasis, and nonmetastasis primary tumor

Figure 5: The level characteristics of CC phylogenetic tree.

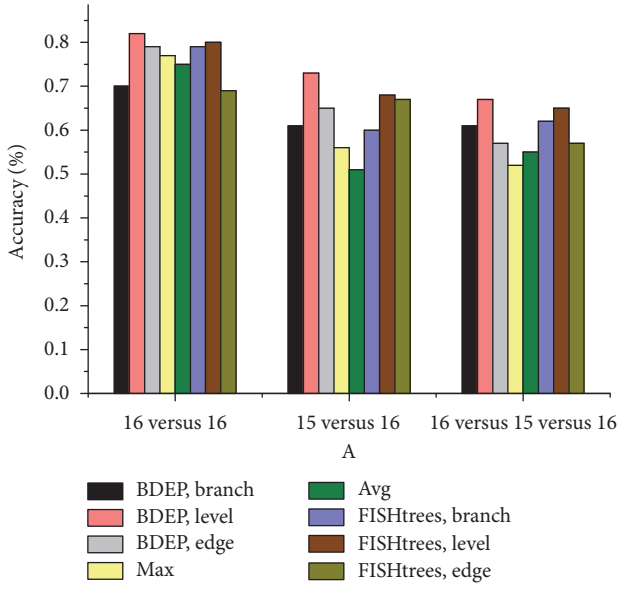


Figure 6: The SVM classification results of different features.

have significant differences for tumors of different states. And the gene under different selection pressures would lead to the different pathways of tumor subcellular expansion. The results on classification experiments show that our tree-based features are in general better than data-based features in distinguishing tumor, which provides more accurate and more comprehensive pathological guidance for clinical diagnosis and treatment. The association between genes is the key point to build and understand tumor progression; combining CNV data with other omics data (RNA and DNA methylation) would be a better strategy for tumor phylogenetic tree inference.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study is supported by the Program for New Century Excellent Talents in University (Grant no. NCET-10-0365), National Natural Science Foundation of China (Grant nos. 11171369, 61272395, 61370171, 61300128, 61472127, 61572178, and 61672214), National Natural Science Foundation of Hunan Province (Grant no. 12JJ2041), and the Planned Science and Technology Project of Hunan Province (Grant nos. 2009FJ3195 and 2012FJ2012).

References

[1] R. Weinberg, The Biology of Cancer. Garland science, 2013.
[2] P. C. Nowell, "The clonal evolution of tumor cell populations," *Science*, vol. 194, no. 4260, pp. 23–28, 1976.
[3] C. Swanton, "Intratumor heterogeneity: Evolution through space and time," *Cancer Research*, vol. 72, no. 19, pp. 4875–4882, 2012.
[4] M. Greaves and C. C. Maley, "Clonal evolution in cancer," *Nature*, vol. 481, no. 7381, pp. 306–313, 2012.
[5] N. Beerenwinkel, R. F. Schwarz, M. Gerstung, and F. Markowetz, "Cancer evolution: Mathematical models and computational inference," *Systematic Biology*, vol. 64, no. 1, pp. e1–e25, 2015.
[6] N. Navin, A. Krasnitz, L. Rodgers et al., "Inferring tumor progression from genomic heterogeneity," *Genome Research*, vol. 20, no. 1, pp. 68–80, 2010.

[7] S. Nik-Zainal, P. Van Loo, D. C. Wedge et al. et al., "The life history of 21 breast cancers," *Cell*, vol. 149, no. 5, pp. 994–1007, 2012.

[8] F. Strino, F. Parisi, M. Micsinai, and Y. Kluger, "TrAp: a tree approach for fingerprinting subclonal tumor composition," *Nucleic Acids Research*, vol. 41, no. 17, p. e165, 2013.

[9] W. Jiao, S. Vembu, A. G. Deshwar, L. Stein, and Q. Morris, "Inferring clonal evolution of tumors from single nucleotide somatic mutations," *BMC Bioinformatics*, vol. 15, no. 1, article no. 35, 2014.

[10] C. A. Miller, B. S. White, N. D. Dees et al., "SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution," *PLoS Computational Biology*, vol. 10, no. 8, Article ID e1003665, 2014.

[11] I. Hajirasouliha, A. Mahmoody, and B. J. Raphael, "A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data," *Bioinformatics*, vol. 30, no. 12, pp. I78–I86, 2014.

[12] M. El-Kebir, L. Oesper, H. Acheson-Field, and B. J. Raphael, "Reconstruction of clonal trees and tumor composition from multi-sample sequencing data," *Bioinformatics*, vol. 31, no. 12, pp. i62–i70, 2015.

[13] V. Popic, R. Salari, I. Hajirasouliha, D. Kashef-Haghighi, R. B. West, and S. Batzoglou, "Fast and scalable inference of multi-sample cancer lineages," *Genome Biology*, vol. 16, no. 1, article no. 91, 2015.

[14] A. Roth, J. Khattra, D. Yap et al., "PyClone: statistical inference of clonal population structure in cancer," *Nature Methods*, vol. 11, no. 4, pp. 396–398, 2014.

[15] R. Redon, S. Ishikawa, K. R. Fitch et al., "Global variation in copy number in the human genome," *Nature*, vol. 444, no. 7118, pp. 444–454, 2006.

[16] B. E. Stranger, M. S. Forrest, M. Dunning et al., "Relative impact of nucleotide and copy number variation on gene phenotypes," *Science*, vol. 315, no. 5813, pp. 848–853, 2007.

[17] A. Subramanian, S. Shackney, and R. Schwartz, "Inference of tumor phylogenies from genomic assays on heterogeneous samples," *Journal of Biomedicine and Biotechnology*, vol. 2012, Article ID 797812, 16 pages, 2012.

[18] L. Oesper, A. Mahmoody, and B. J. Raphael, "THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data," *Genome Biology*, vol. 14, no. 7, article no. R80, 2013.

[19] G. Ha, A. Roth, J. Khattra et al., "TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data," *Genome Research*, vol. 24, no. 11, pp. 1881–1893, 2014.

[20] A. G. Deshwar, S. Vembu, C. K. Yung, G. H. Jang, L. Stein, and Q. Morris, "PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors," *Genome Biology*, vol. 16, no. 1, article no. 35, 2015.

[21] Y. Jiang, Y. Qiu, A. J. Minn, and N. R. Zhang, "Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 113, no. 37, pp. E5528–E5537, 2016.

[22] Y. Li and X. Xie, "Deconvolving tumor purity and ploidy by integrating copy number alterations and loss of heterozygosity," *Bioinformatics*, vol. 30, no. 15, pp. 2121–2129, 2014.

[23] Z. Yu, A. Li, and M. Wang, "CloneCNA: detecting subclonal somatic copy number alterations in heterogeneous tumor samples from whole-exome sequencing data," *BMC Bioinformatics*, vol. 17, no. 1, article no. 310, 2016.

[24] M. El-Kebir, G. Satas, L. Oesper, and B. J. Raphael, "Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures," *Cell Systems*, vol. 3, no. 1, pp. 43–53, 2016.

[25] E. Shapiro, T. Biezuner, and S. Linnarsson, "Single-cell sequencing-based technologies will revolutionize whole-organism science," *Nature Reviews Genetics*, vol. 14, no. 9, pp. 618–630, 2013.

[26] N. Navin, J. Kendall, J. Troge et al., "Tumour evolution inferred by single-cell sequencing," *Nature*, vol. 472, no. 7341, pp. 90–95, 2011.

[27] G. Pennington, C. A. Smith, S. Shackney, and R. Schwartz, "Reconstructing tumor phylogenies from heterogeneous single-cell data," *Journal of Bioinformatics and Computational Biology*, vol. 5, no. 2 A, pp. 407–427, 2007.

[28] S. A. Chowdhury, S. E. Shackney, K. Heselmeyer-Haddad, T. Ried, A. A. Schäffer, and R. Schwartz, "Phylogenetic analysis of multiprobe fluorescence in situ hybridization data from tumor cell populations," *Bioinformatics*, vol. 29, no. 13, pp. i189–i198, 2013.

[29] S. A. Chowdhury, S. E. Shackney, K. Heselmeyer-Haddad, T. Ried, A. A. Schäffer, and R. Schwartz, "Algorithms to model single gene, single chromosome, and whole genome copy number changes jointly in tumor phylogenetics," *PLoS Computational Biology*, vol. 10, no. 7, Article ID e1003740, 2014.

[30] S. A. Chowdhury, E. M. Gertz, D. Wangsa et al., "Inferring models of multiscale copy number evolution for single-tumor phylogenetics," *Bioinformatics*, vol. 31, no. 12, pp. i258–i267, 2015.

[31] E. M. Gertz, S. A. Chowdhury, W.-J. Lee et al., "FISHtrees 3.0: Tumor phylogenetics using a ploidy probe," *PLoS ONE*, vol. 11, no. 6, Article ID e0158569., 2016.

[32] J. Liu, S. Ranka, and T. Kahveci, "Markers improve clustering of CGH data," *Bioinformatics*, vol. 23, no. 4, pp. 450–457, 2007.

[33] R. Shamir, M. Zehavi, and R. Zeira, "A linear-time algorithm for the copy number transformation problem," in *27th Annual Symposium on Combinatorial Pattern Matching (CPM 2016)*, vol. 54 of *Leibniz International Proceedings in Informatics (LIPIcs)*, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.

[34] S. L. Hakimi, "Steiner's problem in graphs and its implications," *Networks*, vol. 1, no. 2, pp. 113–133, 1971.

[35] F. K. Hwang, D. S. Richards, and P. Winter, *The Steiner Tree Problem*, vol. 53, Elsevier, 1992.

[36] Y.-J. Chu and T.-H. Liu, "On shortest arborescence of a directed graph," *Scientia Sinica*, vol. 14, no. 10, p. 1396, 1965.

[37] R. M. Karp, "Reducibility among combinatorial problems," in *Complexity of Computer Computations*, pp. 85–103, Springer, New York, NY, USA, 1972.

[38] R. Storn and K. Price, "Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces," *Journal of Global Optimization*, vol. 11, no. 4, pp. 341–359, 1997.

[39] J. Ilonen, J.-K. Kamarainen, and J. Lampinen, "Differential evolution training algorithm for feed-forward neural networks," *Neural Processing Letters*, vol. 17, no. 1, pp. 93–105, 2003.

[40] R. Joshi and A. C. Sanderson, "Minimal representation multisensor fusion using differential evolution," *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 29, no. 1, pp. 63–76, 1999.

[41] T. Rogalsky, S. Kocabiyik, and R. W. Derksen, "Differential evolution in aerodynamic optimization," *Canadian Aeronautics and Space Journal*, vol. 46, no. 4, pp. 183–190, 2000.

[42] R. Storn, "On the usage of differential evolution for function optimization," in *Proceedings of the Biennial Conference of the North American Fuzzy Information Processing Society (NAFIPS '96)*, pp. 519–523, June 1996.

[43] D. Wangsa, K. Heselmeyer-Haddad, P. Ried et al., "Fluorescence in situ hybridization markers for prediction of cervical lymph node metastases," *The American Journal of Pathology*, vol. 175, no. 6, pp. 2637–2645, 2009.

[44] K. Heselmeyer-Haddad, L. Y. Berroa Garcia, A. Bradley et al., "Single-cell genetic analysis of ductal carcinoma in situ and invasive breast cancer reveals enormous tumor heterogeneity yet conserved genomic imbalances and gain of MYC during progression," *The American Journal of Pathology*, vol. 181, no. 5, pp. 1807–1822, 2012.

[45] H. Kanao, T. Enomoto, T. Kimura et al., "Overexpression of LAMP3/TSC403/DC-LAMP promotes metastasis in uterine cervical cancer," *Cancer Research*, vol. 65, no. 19, pp. 8640–8645, 2005.

[46] K. L. Mine, N. Shulzhenko, A. Yambartsev et al., "Gene network reconstruction reveals cell cycle and antiviral genes as major drivers of cervical cancer," *Nature Communications*, vol. 4, article 1806, 2013.

[47] F. Pedregosa, G. Varoquaux, A. Gramfort et al., "Scikit-learn: machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.