Journal of Information Technology Research
Volume 15 • Issue 1

Integration of Multi-Omics Data to Identify Cancer Biomarkers

Peng Li, School of Artificial Intelligence, Beijing Normal University, China
Bo Sun, School of Artificial Intelligence, Beijing Normal University, China

ABSTRACT

A novel method for integrating multi-omics data, including gene expression, copy number variation, DNA methylation, and miRNA data, is proposed to identify biomarkers of cancer prognosis. First, survival analysis was performed for these four types of omics data to obtain survival-related genes. Next, survival-related genes detected in at least two types of omics data were selected as candidate genes. The four types of omics data only composed of candidate genes were subjected to dimension reduction using an autoencoder to obtain a one-dimensional data representation. The mRMR algorithm was used to screen for key genes. This method was applied to lung squamous cell carcinoma, and 20 cancer-related genes were identified. Gene function analysis revealed that the genes were related to cancer. Using survival analysis, the genes were verified to distinguish between high- and low-risk groups. These results indicate that the genes can be used as biomarkers for cancer.

KEYWORDS

Biomarker, Cancer-Related Genes, Dimension Reduction, Feature Selection, Multi-Omics Data

INTRODUCTION

Cancer, as a complex disease, is not only controlled by individual genes and genetic factors but is also related to the environment and living habits. These factors affect gene expression and thereby influence the occurrence and development of cancer. Biomarkers, such as genes, miRNAs, proteins, metabolites, are biological entities that can determine whether cells, tissues, or individuals are normal or have diseases (Ideker & Sharan, 2008). In the medical field, biomarkers can help diagnose diseases, predict disease development trends, predict the response of patients after treatment, and thus achieve precise and effective treatment for patients. To date, no effective diagnosis and treatment methods have been determined for many types of cancer. Therefore, identifying biomarkers that recognize the early characteristics of cancer and determining the mechanism of cancer occurrence and development are vital.

Traditional cancer biomarkers, such as carcinoembryonic antigens and tumor tissue images, can only detect cancer in the late stages and are not useful for the treatment of patients with cancer. The cure rate and survival rate in patients with cancer are relatively low. Therefore, early detection and timely treatment are necessary to improve these rates.

DOI: 10.4018/JITR.2022010105

The emergence of next-generation sequencing technology has greatly accelerated cancer research. The use of gene expression data to identify cancer-related genes and biomarkers has accelerated the process of individualized treatment (Dancik, 2015). Some studies used gene expression data to distinguish between normal and tumor samples (Nannini et al., 2009). Other studies used gene expression data to detect different states of cancer development (van't Veer et al., 2002; Klahan et al., 2016). However, because gene expression data often include small sample numbers and noise, using only gene expression data limits the discovery of new candidate cancer genes.

In general, gene expression can be regulated by heterogeneous multi-level regulatory factors such as copy number, DNA methylation, transcription factors, and miRNAs (Cancer Genome Atlas Research N, 2012; 2013). High-throughput sequencing can be performed to accurately obtain various biological data at various stages of organism development. These data are collectively referred to as multi-omics data (Reuter et al., 2015) and include multiple types of datasets, such as genomics, epigenomics, transcriptomics, proteomics, metabolomics, and microbiomics data. Using various omics techniques, we are able to understand diseases from a variety of perspectives. Many studies have used DNA methylation, micro RNA (miRNA), protein-protein interaction network (PPIN), or other data to identify cancer-related biomarkers (Zhao et al., 2017; Capper et al., 2018; Liu et al., 2017; Zhou et al., 2016; Wu et al., 2014). However, most methods do not effectively integrate multi-omics data to identify cancer-related genes and biomarkers. Although the use of single-omics data to identify cancer-related genes has yielded many valuable results, a single data source does not provide complete information for a gene, and the results are significantly affected by noise.

In recent years, researchers have used multi-omics data to excavate biomarkers and have made some progress. Moon & Nakai (2018) proposed an integrative analysis of gene expression and DNA methylation using an autoencoder and Welch's t-test to identify candidate cancer biomarkers. Jamal et al. (2016) used a machine learning approach, which integrated topological features from protein-protein interaction networks, to identify candidate Alzheimer's disease-related genes. Jahid & Ruan (2011) integrated gene expression data for breast cancer into the protein-protein interaction networks and identified disease causal genes connecting differentially expressed genes in the network. The results showed that the proposed biomarkers are more stable than those selected by other methods. Cun & Fröhlich (2013) integrated different network information as well as mRNA and miRNA expression data for biomarker discovery. They combined multi-omics data into a classifier, leading to more reproducible and biologically interpretable biomarkers. Increasing studies have used multi-omics data to evaluate gene patterns and the pathogenesis of cancer. Nguyen & Ho (2012) presented a semi-supervised learning method to predict disease-related genes, including cancer-related genes, by integrating genomic and proteomic data. Sanchez-Garcia et al. (2014) integrated many types of genomic data such as gene expression and copy number variation to identify driver genes of breast cancer. Martínez-Ballesteros et al. (2017) presented the integration of three machine learning techniques, including decision trees, quantitative association rules, and hierarchical clustering to analyze Alzheimer's disease genes.

These methods integrate biological data at different levels for different types of diseases. Currently, omics data is widely used to identify molecular biomarkers, but with low reproducibility (Hu et al., 2011). During cancer formation, development, and deterioration, mutations in key genes destabilize the biological network structure, leading to an imbalance in biological systems, driving the whole system towards cancer formation. As key gene variants accumulate, the cancer state is further aggravated. However, among all mutations, few genes drive cancer development. Therefore, many research groups have studied how to integrate multi-omics data to effectively identify biomarkers.

In this study, a novel method for integrating multi-omics data is proposed to identify markers of cancer prognosis. Specifically, survival analysis was performed using four types of omics data, including gene expression, copy number variation, DNA methylation, and miRNA data, to obtain survival-related genes. Then, survival-related genes in at least two types of omics data were selected as candidate genes. Next, the four types of omics data composed only of the candidate genes were

subjected to dimension reduction using an autoencoder to obtain one-dimensional data representation. The maximum correlation relevance minimum redundancy (mRMR) feature selection algorithm was used to screen for key genes. Finally, functional analysis and survival analysis of key genes were performed to identify prognostic biomarkers.

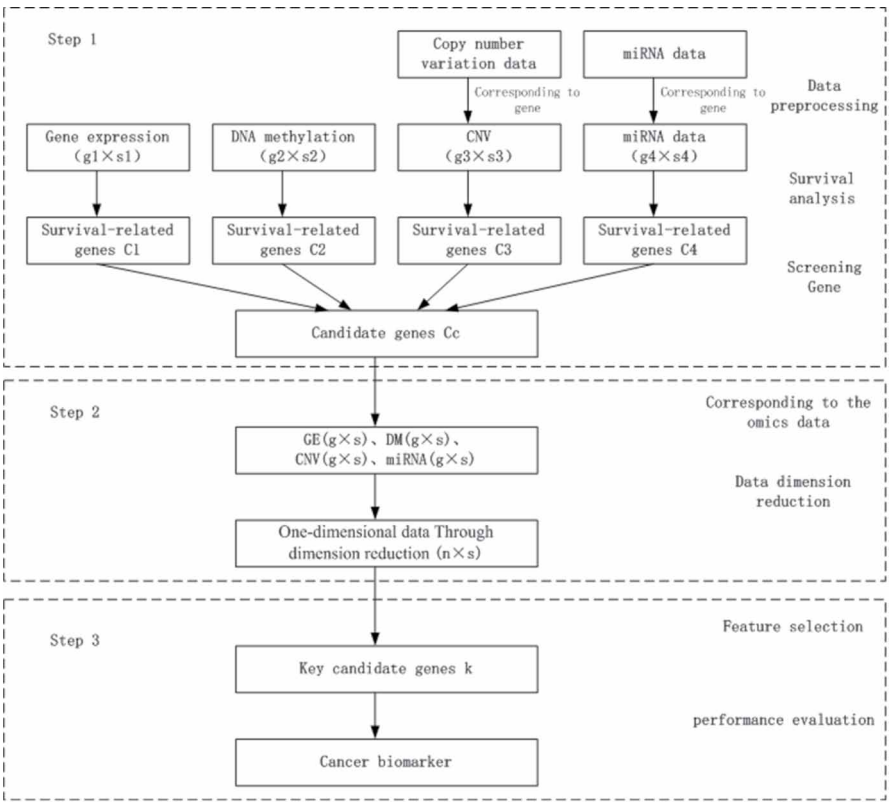
METHODS

Cancer biomarkers, which are useful for the early diagnosis of cancer, have important research value in various fields. Although many valuable results have been achieved in identifying cancer-related genes using only single-omics data, a single data source does not provide complete information about the genes, and using this data for cancer research is greatly affected by noise. Here, multi-omics data were used to identify key genes and biomarkers of cancer. The method proposed in this paper mainly includes three steps, and the flowchart is shown in Figure 1.

Screening Candidate Genes

We integrated gene expression, copy number variation, DNA methylation, and miRNA data to screen for survival-related genes as prognostic markers of cancer. For miRNA expression data and copy number variation data, the features were correspond to the genes respectively. Thus, in the miRNA data matrices and copy number variation data matrices, rows represent genes and columns represent the samples. As the same gene may be connected to multiple miRNAs, the mean values of multiple miRNAs were mapped to the gene.

Figure 1. Flowchart for Identifying Prognosis Biomarkers of Cancer



Survival analysis was performed for each type of omics data. Genes in each omics data were screened by univariant Cox proportional hazard regression model (CPH) to obtain survival-related genes.

If the p-value of the likelihood ratio test of a gene was less than 0.05, the gene was considered as survival-related in the dataset. Thus, survival-related genes C1, C2, C3, and C4 corresponding to gene expression, copy number variation, DNA methylation, and miRNA data were obtained. If a gene was survival-related in at least two types of omics data, this gene was selected and placed in candidate gene set Cc.

Integrating Multi-Omics Data Based on the Autoencoder

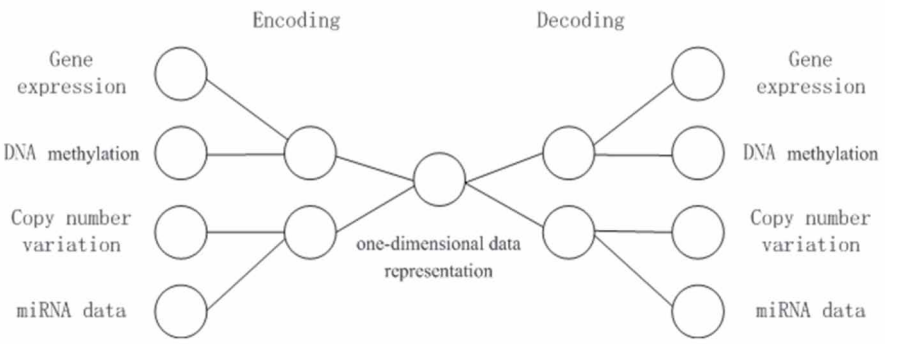
First, according to the candidate genes, the four types of omics data were organized into the matrix with the same genes and same samples, with rows representing candidate genes and columns representing samples. Because of the different scales of the data, the data were normalized to the range [0, 1] using min-max normalization. To improve data processing, 4-bit significant digit was reserved for each standardized data.

An autoencoder was used to reduce the dimension of the four types of omics data to obtain their one-dimensional data representation. The autoencoder is an unsupervised neural network trained to enable copying of its input to its output. Its core function is to learn deeper or higher representations of input data. It has a hidden layer (named as h), which describes the encoding that represents the input. The network has two parts:

Encoding function: $h = f(x)$
Decoding function for reconstruction: $r = g(h)$

When the number of input layer nodes is greater than the number of hidden layer nodes, transformation from the input layer to the hidden layer is essentially a process of dimension reduction, thereby achieving the effect of compressing the input layer. Figure 2 shows the structure of dimension reduction of multi-omics data using an autoencoder. By setting the number of hidden nodes to 1, the four-dimensional data can be reduced to one-dimensional data. To reduce the model size, the weights of the encoder matrix were bundled with the weights of the decoder matrix so that the decoder weight matrix was the transpose of the encoder weight matrix. f and g were the activation functions of the encoder and decoder, respectively, and the ReLU activation function was used for nonlinear transformation.

Figure 2. Structure of the autoencoder applied for dimension reduction



Identification of Key Genes as Prognosis Biomarkers

The mRMR algorithm was applied to candidate genes for feature selection to screen for key genes. The mRMR algorithm (Ding & Peng, 2005) is a commonly used feature selection algorithm, and widely used in the bioinformatics field to classify feature sets by calculating the correlation between sample tags and features.

The mRMR algorithm was applied to gene expression data to obtain key candidate genes from candidate genes. The sample tag includes two types of survival states of the patients, which are dead or alive. The mRMR algorithm requires the feature set to be a discretized value. The expression values of the candidate genes were discretized using the mean μ and variance σ calculated based on the candidate genes expression values. If the expression value was greater than $(\mu+\sigma/2)$, it was converted to 1; if the expression value was less than $(\mu-\sigma/2)$, it was converted to -1; if the expression value was between $(\mu-\sigma/2)$ and $(\mu+\sigma/2)$, it was converted to 0. These three states correspond to over-expression (1), low-expression (-1), and normal expression (0). The top 20 candidate genes were selected as key genes.

For this study, the CPH and the Kaplan–Meier survival analysis were implemented with the use of R package survival. Tensorflow 1.8 was utilized for the implementation of autoencoder. All data pre-processing, performance measurements and PCA dimension reduction were implemented using Python and Scikit-learn 0.19.2. Hierarchical clustered heatmaps was implemented with the use of R package pheatmap.

MATERIALS AND RESULTS ANALYSIS

Materials

Lung squamous cell carcinoma (LUSC) is a common type of lung cancer, but its mechanism of pathogenesis is unclear. In the paper, taking LUSC as the object, the identifying method for prognosis biomarkers of cancer based on multi-omics data was applied.

The multi-omics data used in the paper was downloaded from Firebrowse, as shown in Table 1, which describes the number of samples and factors including gene expression, copy number variation, DNA methylation, miRNA expression data. In addition, it also includes clinical data that provides information on the survival time, age, and gender of LUSC patients.

Analysis of Experimental Results

Analysis of Gene Function

Using mRMR feature selection algorithm, 20 key genes closely related to survival were obtained from the candidate survival genes, namely GOLGA8A, ZNF665, CNGA1, TM6SF1, LMTK3, NACC1, TROVE2, PACSIN2, MYLIP, TAOK2, RAD52, AIFM3, CLK1, NEK6, CDC42BPG, TNK2, DYNC1I2, SLC36A4, SOHLH2, and GALC. The chromosomal information of each key gene, including the chromosome number of the gene and the start site information of the chromosome to which the gene belongs, is shown in Table 2.

Table 1. Multi-omics data

Name	The number of samples	The number of factors (genes)
Gene expression data	552	20531
Copy number variation	1035	19340
DNA Methylation	412	21053
miRNA expression	523	13047

Table 2. Chromosome information of key genes

Gene name	Gene ID	Chromosome	Start site	End site
GOLGA8A	23015	hs15	34379068	34437466
ZNF665	79788	hs19	53163299	53193425
CNGA1	1259	hs4	47935015	48016718
TM6SF1	53346	hs15	83107486	83145403
LMTK3	114783	hs19	48485271	48513926
NACC1	112939	hs19	13116848	13141147
TROVE2	6738	hs1	193059422	193091777
PACSIN2	11252	hs22	42869766	43016174
MYLIP	29116	hs6	16129086	16151015
TAOK2	9344	hs16	29973867	29992261
RAD52	5893	hs12	911028	991195
AIFM3	150209	hs22	20965130	20981360
CLK1	1195	hs2	200853009	200864744
NEK6	10783	hs9	124257606	124352442
CDC42BPG	55561	hs11	64823809	64844686
TNK2	10188	hs3	195863364	195909009
DYNC1I2	1781	hs2	171687409	171750158
SLC36A4	120103	hs11	93144171	93386038
SOHLH2	54937	hs13	36168208	36214615
GALC	2581	hs14	87933014	87993665

Among the 20 key genes a total of 18 genes related to cancer or serious diseases could be found in GeneCards. The information of the key genes is shown in Table 3.

Distinguish Between Normal Samples and Tumor Samples

To verify the important impact of the genes on development of LUSC, gene expression data was downloaded from Firebrowse for distinguishing between normal samples and tumor samples. The dataset has a total of 552 samples containing 501 LUSC tumor samples and 51 normal samples. The performance of the proposed method was measured by the classification performance according to gene expression data of these 20 key genes. First, we took into account the 20 keys genes from the dataset and calculated four performance measures, i.e. the accuracy, precision, recall, and F1 score with five classifiers, i.e. support vector machine (SVM), logistic regression, naïve bayes, decision tree, and random forest. In addition, we performed hierarchical clustering for the gene expression data of the 20 key genes, and used heatmaps to reveal classification results.

10-fold cross-validation test was conducted for performance evaluation. The 20 key genes were chosen from the gene expression gene dataset for the test. Five classifiers have been utilized to calculate the four performance measures. Table 4 demonstrates the detailed results of the performance evaluation. As described in the table 4, these 20 genes can distinguish remarkably between tumor samples and normal samples well.

Table 3. Information of cancer-related key genes

Gene	Description	Gifts	Score
GOLGA8A	Golgin A8 Family Member A	35	16.05
ZNF665	Zinc Finger Protein 665	38	16.05
CNGA1	Cyclic Nucleotide Gated Channel Alpha 1	51	16.05
TM6SF1	Transmembrane 6 Superfamily Member 1	38	16.05
LMTK3	Lemur Tyrosine Kinase 3	39	16.05
NACC1	Nucleus Accumbens Associated 1	46	16.05
TROVE2	Ro60, Y RNA Binding Protein	35	16.05
PACSIN2	Protein Kinase C And Casein Kinase Substrate In Neurons 2	45	16.05
MYLIP	Myosin Regulatory Light Chain Interacting Protein	45	16.05
TAOK2	TAO Kinase 2	46	16.05
RAD52	RAD52 Homolog, DNA Repair Protein	48	16.05
AIFM3	Apoptosis Inducing Factor Mitochondria Associated 3	42	16.05
CLK1	CDC Like Kinase 1	48	16.05
NEK6	NIMA Related Kinase 6	46	16.05
CDC42BPG	CDC42 Binding Protein Kinase Gamma	43	16.05
TNK2	Tyrosine Kinase Non Receptor 2	50	16.05
DYNC1I2	Dynein Cytoplasmic 1 Intermediate Chain 2	43	16.05
SLC36A4	Solute Carrier Family 36 Member 4	39	16.05
SOHLH2	Spermatogenesis And Oogenesis Specific Basic Helix-Loop-Helix 2	37	16.05
GALC	Galactosylceramidase	50	16.05

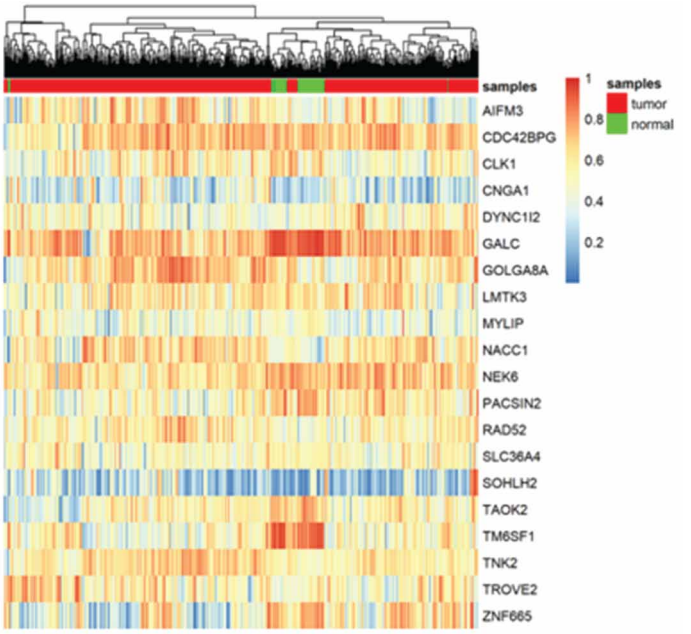
Table 4. Distinguishing normal samples and tumor samples

Performance measures	SVM	Logistic regression	Naive Bayes	Decision tree	Random forest
Accuracy	98.4%	98.2%	90.8%	97.1%	99.5%
Precision	98.4%	98.2%	90.8%	98.2%	99.6%
Recall	99.8%	99.8%	100%	98.6%	99.8%
F1 score	99.1%	98.9%	95.1%	98.4%	99.7%

We performed hierarchical clustering for the gene expression data of the 20 key genes. Euclidean distance was used as distance metric. The clustered heatmaps is shown in Figure 3.

To the first row of the figure, the red represents tumor samples and the green represents normal samples. Each row of other parts represents a gene, and each column represents a sample, indicating expression value of a gene under a specific sample. The red and blue color indicates the size of the gene expression value. According to Figure 3, it could be also found that these 20 key genes could distinguish obviously between tumor samples and normal samples.

Figure 3. Distinguishing between normal samples and tumor samples



Survival Analysis

To demonstrate important roles of the key genes in the development and progression of LUSC, LUSC dataset was downloaded from TCGA for survival analysis. The dataset has a total of 205 LUSC samples containing 103 low-risk group tumor samples and 102 high-risk group tumor samples.

To validate the effectiveness of the proposed method, the method proposed in this paper was compared with PCA dimension reduction, single-omics data including gene expression data and DNA methylation data respectively. We obtained key genes using principal component analysis (PCA) dimension reduction, gene expression data and DNA methylation data, respectively. For gene expression data or DNA methylation data, we first utilized univariant Cox proportional hazard regression model to screen candidate genes, and then used mRMR feature selection method to obtain 20 key genes.

We performed survival analysis on these four groups of key genes using the Cox proportional hazards model. The Kaplan–Meier curves of the analysis results are shown in Figure 4-7.

The x-axis represents time (unit: month), and y-axis represents global survival ratio. The red line represents the high-risk group and the green line represents the low-risk group. In the upper right corner, the number on the left represents the number of people in each group. The number with the '+' sign in the middle represents the number of lost visitors, and the number on the right represents the concordance index (c-index).

By comparing the two curves, we found that differences gradually increased over time, and all key genes could significantly distinguish between the two groups of patients ($p < 0.05$) with respect to survival. This illustrates the effectiveness of the framework we propose. From the figures, we can see that p-value (p=9.794e-07) of Figure 4 is the lowest. So, compared with the key genes using PCA dimension reduction method, gene expression data and DNA methylation data to screen, the proposed method has the best performance. Because PCA dimension reduction is a linear transformation, using autoencoder dimension reduction can achieve better performance than PCA dimension reduction. The regulation of genes is a complex process, and gene expression

Figure 4. Survival analysis for the key genes using proposed method

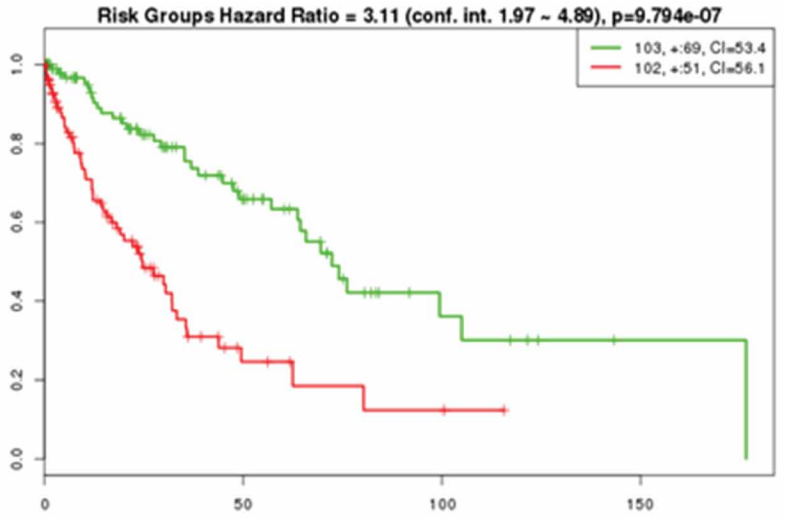
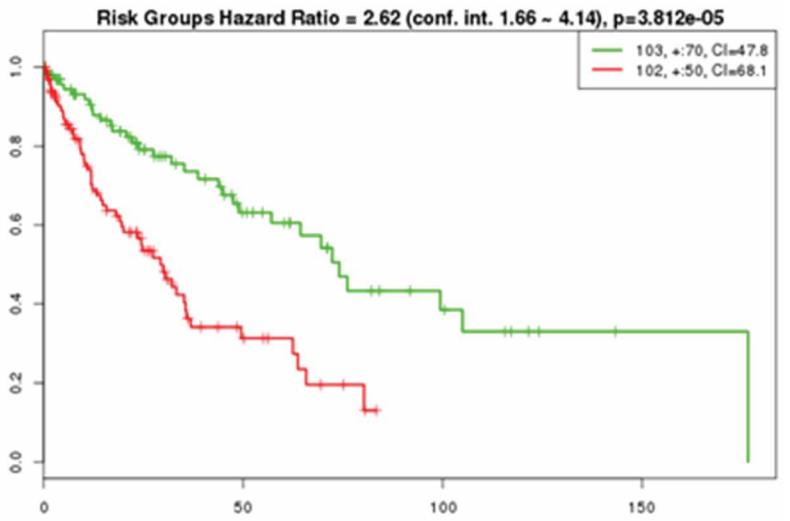


Figure 5. Survival analysis for the key genes using PCA dimension reduction to integrating multi-omics data



can be controlled by a variety of regulatory factors. So, integrating multi-omics data has better performance than using only a single-multi data.

DISCUSSIONS

According to the paper "Cancer statistics, 2018", published by the authoritative journal "CA: A Cancer Journal for clinicians", there will be 18.1 million new cancer cases and 9.6 million cancer deaths worldwide in 2018, and the morbidity rate and mortality rate are increasing year by year (Siegel et al., 2018). Among them, lung cancer, breast cancer and colorectal cancer are the three most common cancers in the world, and with the mortality rates being first, fifth and second respectively. In recent

Figure 6. Survival analysis for the key genes using gene expression data

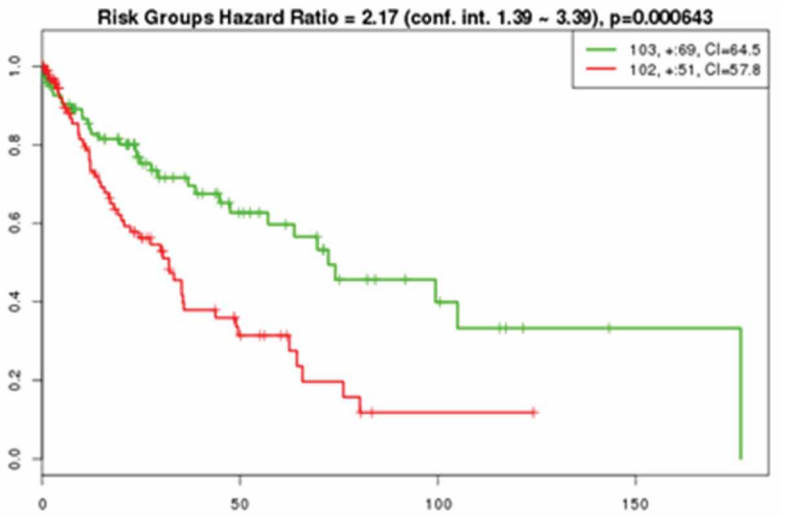
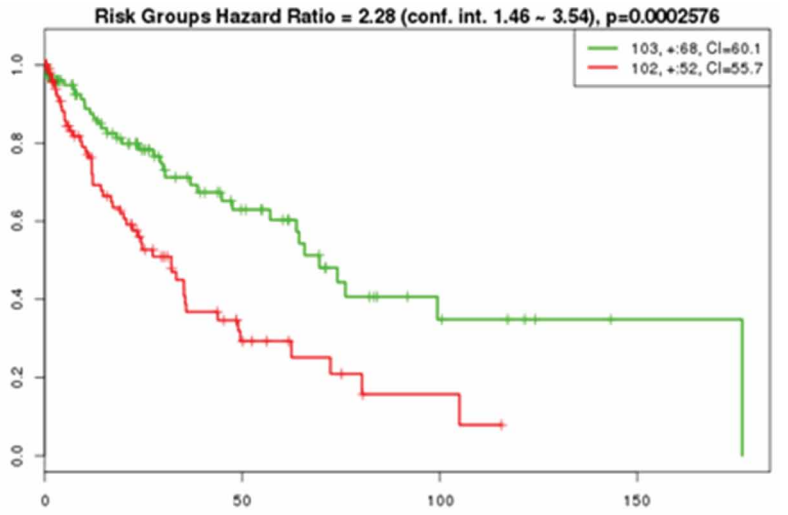


Figure 7. Survival analysis for the key genes using DNA methylation data



50 years, the incidence rate and mortality rate of lung cancer have increased significantly in many countries The cause of lung cancer is still not fully understood. Among various types of lung cancer, non-small cell lung cancer (NSCLC) accounts for about 85% of lung cancer. NSCLC mainly includes lung squamous cell carcinoma (LUSC), lung adenocarcinoma (LUAD) and lung large cell carcinoma (LULC), and LUSC accounts for about 50% of NSCLC. At present, many research groups all over the world are studying lung cancer.

In this paper, a new mining method of cancer biomarkers based on multi-oimcs data is proposed. Through this method, we get key genes composed of 20 genes. The functions and pathways of these genes are closely related to the occurrence and development of cancer. Through survival analysis, it is found that the genes can distinguish the high and low risk group of the patient, and have good prognosis performance.

In recent years, more and more studies have used multi-omics data to study cancer in order to more accurately understand the mechanism of cancer occurrence. According to different using methods, integration methods of multi-omics data can be divided into many methods, such as feature concatenation-based method, Bayesian-based method, dimension reduction-based method, and network-based method. Each of these methods has its own advantages and disadvantages, and can be selected according to its characteristics in future research.

Multi-omics data are usually high-dimensional and have the following characteristics: (1) The number of features in multi-omics data is much larger than the number of samples. (2) There is noise in each dataset. (3) There is information relevance among the multi-omics data.

In this paper, dimensionality reduction-based method was used to integrate multi-omics data. Dimensionality reduction-based method is the most effective way to solve clustering and classification analysis of high-dimensional data. High-dimensional multi-omics data can be projected into a low-dimensional subspace containing some major biological processes through dimensionality reduction. Matrix decomposition methods, such as PCA, independent component analysis (ICA) and non-negative matrix factorization (NMF) are commonly used data dimensionality reduction methods. In recent years, the deep learning framework has been applied in various fields and achieved excellent results. Deep learning framework is different from traditional machine learning methods in how representations are learned from the raw data. Other dimensionality reduction methods than matrix decompositions, deep learning framework, such as autoencoder, restricted Boltzmann machine (RBM), and deep belief network (DBN), can be applied as well.

Using literature validation, each gene of the cancer biomarkers identified by the proposed method is analyzed. Kim et al. (2013) investigated whether dose-dependent alteration in gene expression is implicated in clinical outcomes of lung cancer, and found GOLGA8A was clinical association of Up-pattern genes regulated by the Ethanol extract of the seeds of Descurainia sophia (EEDS) with survival from lung cancer. CpG island methylator phenotype (CIMP) is a major mechanism for colorectal cancer. By comparing the CIMP+ and CIMP- colorectal cancer samples with mRMR and incremental feature selection (IFS) methods, Zhang et al. (2018) found ZNF665 was highly expressed in CIMP- patients. Olfactory receptors (OR) activation has been demonstrated to have influence on cancer cell growth and progression. Weber et al. (2017) analyzed the RNA-Seq data of the cell line with a focus on the signaling cascade components, and CNGA1 show low expression rates. Using significance analysis of microarrays, Zhang et al. (2017) identified TM6SF1 as differentially expressed genes (DEG) between LUSC and normal controls. In contrast to the normal lung tissues, increased LMTK3 expression was found in the NSCLC tissues, and was mainly located on the cytoplasm and the nuclei of cancer cells (Zhang et al, 2015). The rising incidence of oral tongue squamous cell carcinoma (OTSCC) in patients prompted researches to develop a new cell line. NACC1 expression was a molecular biomarker for OTSCC (Wang et al, 2017). TROVE2 was located at gene expression pathway hsa05322 significantly associated with consortium lung phenotype (Deepika et al, 2018). Monotonically expressed genes (MEGs) are genes whose expression values increase or decrease monotonically as a disease advances or time proceeds. PACSIN2 was a monotonically expressed gene in the ascending order across the risk levels of death in NSCLC patients (Tian, 2019). Low level of MYLIP was associated with poor survival in patients with lung cancer (Xue et al., 2017). TAOK2 exhibited deregulation pattern, and was associated with the NSCLC developmental process (Guo et al., 2015). Recent genome-wide association studies show that RAD52, that is associated with increased lung cancer risk, is significantly associated with the development of LUSC. Somatic overexpression of RAD52 was confirmed to be significant in LUSC tumors (Lieberman et al., 2016). Zheng et al. (2019) investigated the crystal structure, clinical and prognostic implications of AIFM3 in breast cancer (BC). AIFM3 was significantly more expressed in BC tissues than in normal tissues, might be a potential biomarker for predicting prognosis in BC. Disruption of CLK1 causes pleiotropic cell cycle defects and loss of proliferation, whereas CLK1 over-expression is associated with various cancers (Dominguez et al., 2016). NEK6 was overexpressed in a subset of human prostate cancers

(Choudhury et al., 2017). The mammalian variation analogous to G723S in CDC42BPG was detected in LUAD (Ferguson et al., 2015). EGFR-dependent cell migration plays an important role in lung cancer progression. NEDD4 is collaborated with TNK2 to regulate transport of the EGFR-loaded endosomes to MVBs/lysosomes (Shao et al., 2018). The present study examined the interaction between celecoxib and sorafenib in two human liver tumor cell lines HepG2 and Huh7. DYNC1I2 was differentially expressed in HepG2 and Huh7 cells upon combined sorafenib+celecoxib treatment (Cervello et al., 2013). Androgen steroid hormones are key drivers of prostate cancer. SLC36A4 regulated by androgen was identified in prostate cancer cells (Munkley et al., 2018). SOHLH2 was not only identified as a tumor suppressor in the pathogenesis of ovarian cancer, but also observed expression downregulated in the metastatic breast cancer (Ji et al., 2016). Peng et al. (2015) performed methylation and expression analysis on GALC gene in a panel of lung cancer cell lines, found GALC was related to human cancer.

Thus, all these 20 key genes are associated with the occurrence and development of cancer, of which 13 genes are associated closely with lung cancer. The result proves to some extent the effectiveness of our method. Through the functions and pathways analysis of these 20 genes, we found that they were highly correlated with cancer. Through functional analysis and survival analysis, we believe that each of these 20 genes is not only related to cancer survival at its corresponding level, they are also closely related to cancer survival as a whole and can be used as a biomarker. The number of potential biomarkers obtained in the study is very small, which provides great convenience for identification. Excavated genes can help researchers design treatments and early diagnosis of cancer.

CONCLUSION

In the field of life sciences, biomarkers helpful for identifying early signs of cancer and understanding the biological regulatory mechanisms of the occurrence and development of cancer have been evaluated worldwide. As genetic studies become more detailed, increasing evidence indicates that gene expression is affected by multiple levels, and interactions between various regulatory factors may occur. Researchers have designed various mining methods for cancer-related genes and successfully explored a large number of potential cancer-related genes. Among these, integration of multi-omics data for mining cancer gene method has become a research hotspot in recent years.

The emergence of next-generation sequencing technology has greatly accelerated cancer research, providing a foundation for the discovery of important genes related to cancer and determining relationships between these genes. This method integrates multi-omics data from the perspective of data integration, comprehensively considers the factors contributing to cancer development and their interactions, explores the genes that cause cancer, and identifies prognostic biomarkers. This method promotes the application of machine learning and data mining in bioinformatics. In future studies, we will develop a survival analysis model to predict and analyze the survival of patients with cancer.

ACKNOWLEDGMENT

This research was supported by the national natural science foundation of China [grant number 61532014].

REFERENCES

Cancer Genome Atlas Research. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, *490*(7418), 61–70.

Capper, D., Jones, D. T. W., Sill, M., & Hovestadt, V. (2018). DNA methylation-based classification of central nervous system tumours. *Nature*, *555*, 469–474.

Cervello, M., Bachvarov, D., Lampiasi, N., Cusimano, A., Azzolina, A., McCubrey, J. A., & Montalto, G. (2013). Novel Combination of Sorafenib and Celecoxib Provides Synergistic Anti-Proliferative and Pro-Apoptotic Effects in Human Liver Cancer Cells. *PLoS One*, *8*(6), e65569.

Choudhury, A. D., Schinzel, A. C., Cotter, M. B., Lis, R. T., Labella, K., Lock, Y. J., Izzo, F., & Guney, I. (2017). Castration resistance in prostate cancer is mediated by the kinase NEK6. *Cancer Research*, *77*(3), 753–765.

Cun, Y., & Fröhlich, H. (2013). Network and data integration for biomarker signature discovery via network smoothed t-statistics. *PLoS One*, *8*(9).

Dancik, G. M. (2015). An online tool for evaluating diagnostic and prognostic gene expression biomarkers in bladder cancer. *BMC Urology*, *15*, 59.

Ding, C., & Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*, *3*(2), 185–205.

Dominguez, D., Tsai, Y. H., Weatheritt, R., Wang, Y., Blencowe, B. J., & Wang, Z. F. (2016). An extensive program of periodic alternative splicing linked to cell cycle progression. *eLife*, *5*, 25.

Ferguson, B. D., Tan, Y. H. C., Kanteti, R. S., Liu, R., Gayed, M. J., & Everett, E. (2015). Novel EPHB4 Receptor Tyrosine Kinase Mutations and Kinomic Pathway Analysis in Lung Cancer. *Scientific Reports*, *5*, 10641.

Guo, W., Xie, L., Zhao, L., & Zhao, Y. H. (2015). mRNA and microRNA expression profiles of radioresistant NCI-H520 non-small cell lung cancer cells. *Molecular Medicine Reports*, *12*(2), 1857–1867.

Hu, Z.Z., Huang, H., Wu, C.H., Jung, M., Dritschilo, A., & Riegel, A.T. (2011). Omicsbased molecular target and biomarker identification. *Bioinformatics for Omics Data: Methods and Protocols,* 547-571.

Ideker, T., & Sharan, R. (2008). Protein networks in disease. *Genome Research*, *18*(4), 644–652. doi:10.1101/gr.071852.107 PMID:18381899

Kandoth, C., & Schultz, N. (2013). Integrated genomic characterization of endometrial carcinoma. *Nature*, *497*(7447), 67–73.

Jahid, M. J., & Ruan, J. (2011). Identification of biomarkers in breast cancer metastasis by integrating protein-protein interaction network and gene expression data. In *2011 IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS)*. IEEE.

Jamal, S., Goyal, S., & Shanker, A. (2016). Integrating network, sequence and functional features using machine learning approaches towards identification of novel Alzheimer genes. *BMC Genomics*, *17*(1), 807.

Ji, S. F., Zhang, W. F., Zhang, X. L., Hao, C. Y., Hao, A. J., Gao, Q., Zhang, H. Y., Sun, J. H., & Hao, J. (2016). Sohlh2 suppresses epithelial to mesenchymal transition in breast cancer via downregulation of IL-8. *Oncotarget*, *7*(31), 49411–49424.

Kim, B. Y., Lee, J., Park, S. J., Bang, O. S., & Kim, N. S. (2013). Gene expression profile of the A549 human non-small cell lung carcinoma cell line following treatment with the seeds of descurainia sophia, a potential anticancer drug. *Evidence-Based Complementary and Alternative Medicine*, 1–13.

Klahan, S., Huang, W. C., Chang, C. M., Wong, H. S. C., Huang, C. C., Wu, M. S., Lin, Y. C., Lu, H. F., Hou, M. F., & Chang, W. C. (2016). Gene expression profiling combined with functional analysis identify integrin beta1 (ITGB1) as a potential prognosis biomarker in triple negative breast cancer. *Pharmacological Research*, *104*, 31–37.

Lieberman, R., Xiong, D. H., James, M., Han, Y. H., Amos, C. I., Wang, L., & You, M. (2016). Functional characterization of RAD52 as a lung cancer susceptibility gene in the 12p13.33 locus. *Molecular Carcinogenesis*, *55*(5), 953–963.

Liu, Y. S., Zeng, X. X., He, Z. Y., & Zou, Q. (2017). Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *14*(4), 905–915.

Martínez-Ballesteros, M., García-Heredia, J. M., Nepomuceno-Chamorro, I. A., & Riquelme-Santos, J. C. (2017). Machine learning techniques to discover genes with potential prognosis role in Alzheimer's disease using different biological sources. *Information Fusion*, *36*, 114–129.

Moon, M., & Nakai, K. (2018). Integrative analysis of gene expression and DNA methylation using unsupervised feature extraction for detecting candidate cancer biomarkers. *Journal of Bioinformatics and Computational Biology*, *16*(3), 1–20.

Munkley, J., Maia, T. M., Ibarluzea, N., Livermore, K. E., Vodak, D., Ehrmann, I., & James, K. (2018). Androgen-dependent alternative mRNA isoform expression in prostate cancer cells. *F1000 Research*, *7*, 1189.

Nannini, M., Pantaleo, M. A., & Maleddu, A. (2009). Gene expression profiling in colorectal cancer using microarray technologies: Results and perspectives. *Cancer Treatment Reviews*, *35*, 201–209.

Nguyen, T. P., & Ho, T. B. (2012). Detecting disease genes based on semi-supervised learning and protein-protein interaction networks. *Artificial Intelligence in Medicine*, *54*, 63–71.

Peng, J. Z., Chen, B. S., Shen, Z. J., Deng, H. R., Liu, D. G., Xie, X., Gan, X. F., Xu, X., Huang, Z. Q., & Chen, J. (2015). DNA promoter hypermethylation contributes to down-regulation of galactocerebrosidase gene in lung and head and neck cancers. *International Journal of Clinical and Experimental Pathology*, *8*(9), 11042–11050.

Polineni, D., Dang, H., Gallins, P. J., Jones, L. C., Pace, R. G., Stonebraker, J. R., & Leah, A. et al.. (2018). Airway Mucosal Host Defense Is Key to Genomic Regulation of Cystic Fibrosis Lung Disease Severity. *American Journal of Respiratory and Critical Care Medicine*, *197*(1), 79–93.

Reuter, J. A., Spacek, D. V., & Snyder, M. P. (2015). High-throughput sequencing technologies. *Molecular Cell*, *58*, 586–597.

Sanchez-Garcia, F., Villagrasa, P., & Matsui, J. (2014). Integration of genomic data enables selective discovery of breast cancer drivers. *Cell*, *159*(6), 1461–1475.

Shao, G. B., Wang, R. R., Sun, A. Q., Wei, J., Peng, K., Dai, Q., Yang, W. N., & Lin, Q. (2018). The E3 ubiquitin ligase NEDD4 mediates cell migration signaling of EGFR in lung cancer cells. *Molecular Cancer*, *17*, 24.

Siegel, R. L., Miller, K. D., & Jemal, A. (2018). Cancer Statistics, 2018. *CA: a Cancer Journal for Clinicians*, *68*(1), 7–30.

Tian, S. Y. (2019). Identification of monotonically differentially expressed genes for non-small cell lung cancer. *BMC Bioinformatics*, *20*, 177.

van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., & Peterse, H. L. et al.. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, *415*, 530–536.

Wang, S. J., Asthana, S., van Zante, A., Heaton, C. M., Phuchareon, J., Stein, L., & Higuchi, S. et al.. (2017). Establishment and characterization of an oral tongue squamous cell carcinoma cell line from a never-smoking patient. *Oral Oncology*, *69*, 1–10.

Weber, L., Al-Refae, K., Ebbert, J., Jägers, P., Altmüller, J., Becker, C., Hahn, S., Gisselmann, G., & Hatt, H. (2017). Activation of odorant receptor in colorectal cancer cells leads to inhibition of cell proliferation and apoptosis. *PLoS One*, *12*(3), 1–27.

Wu, S. Y., Shao, F. J., & Sun, R. C. (2014). Analysis of human genes with protein-protein interaction network for detecting disease genes. *Physica A*, *398*, 217–228.

Xue, W. H., Li, L. F., Tian, X., Fan, Z. R., Yue, Y., Zhang, C. Q., Ding, X. F., Song, X. Q., & Ma, B. J. et al.. (2017). Integrated analysis profiles of long non-coding RNAs reveal potential biomarkers of drug resistance in lung cancer. *Oncotarget*, *8*(38), 62868–62879.

Zhang, F., Chen, X., Wei, K., Liu, D. M., Xu, X. D., Zhang, X., & Shi, H. (2017). Identification of Key Transcription Factors Associated with Lung Squamous Cell Carcinoma. *Medical Science Monitor*, *23*, 172–206.

Zhang, K. X., Chen, L. J., Deng, H. F., Zou, Y. Y., Liu, J., Shi, H. B., Xu, B., Lu, M. Y., Li, C., Jiang, J. T., & Wang, Z. G. (2015). Serum lemur tyrosine kinase-3: A novel biomarker for screening primary non-small cell lung cancer and predicting cancer progression. *International Journal of Clinical and Experimental Pathology*, *8*(1), 629–635.

Zhang, T. M., Huang, T., & Wang, R. F. (2018). Cross talk of chromosome instability, CpG island methylator phenotype and mismatch repair in colorectal cancer. *Oncology Letters*, *16*(2), 1736–1746.

Zhao, S., Geybels, M. S., & Leonardson, A. (2017). Epigenome-Wide Tumor DNA Methylation Profiling Identifies Novel Prognostic Biomarkers of Metastatic-Lethal Progression in Men Diagnosed with Clinically Localized Prostate Cancer. *Clinical Cancer Research*, *23*, 311–319.

Zheng, A., Zhang, L., Song, X., Wang, Y., Wei, M., & Jin, F. (2019). Clinical implications of a novel prognostic factor AIFM3 in breast cancer patients. *BMC Cancer*, *19*(1), 451.

Zhou, M., Xu, W., & Yue, X. (2016). Relapse-related long non-coding RNA signature to improve prognosis prediction of lung adenocarcinoma. *Oncotarget*, *7*, 29720–29738.

Peng Li received his B.S. in computer science from QingDao University, ShangDong, China, in 1997, and received his M.S. degree in computer science from Shandong University of Technology, China, in 2000. Since 2003, he has been a faculty member as a Lecturer in Beijing University of Civil Engineering and Architecture. He is currently working toward the PhD degree at Beijing Normal University (BNU). His research focuses on machine learning, bioinformatics.

Bo Sun received the B.S. degree in computer science from Beihang University (BUAA), Beijing, China, in 1988, the M.S. and Ph.D. degrees separately in natural language process and computer-aided education from Beijing Normal University (BNU), Beijing, in 1991 and 2003. From 1999 to 2004, he was an Associate Professor with the Computer Science Department at BNU. Since 2004, he has been a Professor with the College of Information Science and Technology, BNU. He has led and accomplished a number of research projects, published more than 90 papers in recent ten years, such as "Affect recognition from facial movements and body gestures by hierarchical deep spatio-temporal features and fusion strategy" in Neural Network, "BNU-LSVED 2.0: Spontaneous multimodal student affect database with multi-dimensional labels", in Signal Processing: Image Communication, etc. His research interests include machine learning, deep neural networks, pattern recognition and natural language processing. Dr. Sun is now a professor of the School of Artificial Intelligence, the director of Intelligent Computing and Software Research Center, at BNU, a senior member of ACM, China Computer Federation and China Society of Image and Graphics.