SCIENTIFIC REPORTS

OPEN

Received: 08 April 2016
Accepted: 08 June 2016
Published: 28 June 2016

A new strategy for exploring the hierarchical structure of cancers by adaptively partitioning functional modules from gene expression network

Junmei Xu[1], Runyu Jing[1], Yuan Liu[1], Yongcheng Dong[2], Zhining Wen[1] & Menglong Li[1]

The interactions among the genes within a disease are helpful for better understanding the hierarchical structure of the complex biological system of it. Most of the current methodologies need the information of known interactions between genes or proteins to create the network connections. However, these methods meet the limitations in clinical cancer researches because different cancers not only share the common interactions among the genes but also own their specific interactions distinguished from each other. Moreover, it is still difficult to decide the boundaries of the sub-networks. Therefore, we proposed a strategy to construct a gene network by using the sparse inverse covariance matrix of gene expression data, and divide it into a series of functional modules by an adaptive partition algorithm. The strategy was validated by using the microarray data of three cancers and the RNA-sequencing data of glioblastoma. The different modules in the network exhibited specific functions in cancers progression. Moreover, based on the gene expression profiles in the modules, the risk of death was well predicted in the clustering analysis and the binary classification, indicating that our strategy can be benefit for investigating the cancer mechanisms and promoting the clinical applications of network-based methodologies in cancer researches.

Researches on analyzing the functions of individual gene have contributed to finding potential oncogenes involved in the initiation and progression of tumors that may be potential targets of anticancer drugs and can be applied to the prognosis of patients[1–8]. Although the biological studies of individual gene play a vital role in clinical treatment, the emergent functions and behaviors of individual genes in the context of the whole biological systems still require further investigation[9,10]. As the biological function is usually completed by the cooperation of a set of genes, the gene network is one of the suitable tools for better understanding the molecular mechanisms of biological processes in the whole systems[11].

The network has been used to represent the interactions or regulations among the elements in the system[12]. The framework of gene network, for which the nodes represent genes and the edges represent the interactions between genes, allows us to explore the integrated functions of the sub-networks as well as the property of the whole network, which serves as a powerful paradigm for genes analysis and provides a system-level understanding of gene functions[13]. Generally, the connections among genes or proteins are created mainly based on the known interactions in the knowledge-based databases, such as Human Protein Reference Database (HPRD)[14], STRING[15], and Biomolecular Interaction Network Database (BIND)[16]. A number of algorithms have been developed to partition the whole network into several sub-networks in order that people can investigate the potential relationships among the sub-networks, and subsequently infer the molecular mechanisms of the complex biological systems[17–30]. Recently, Ghiassian *et al.* successfully extracted the disease modules by systematically analyzing the protein-protein interaction networks of 70 diseases and discovered the pathobiological relationships among

[1]College of Chemistry, Sichuan University, Chengdu, China. [2]College of Life Sciences, Sichuan University, Chengdu, China. Correspondence and requests for materials should be addressed to Z.W. (email: w_zhining@163.com) or M.L. (email: liml@scu.edu.cn)

these diseases with the identified modules[31]. Menche *et al.* further elucidated the pathobiological relationships among the diseases based on the similarities of the disease modules[32].

In clinical cancer researches, the gene networks are also benefit for exploring the hierarchical structure of the biological system of a complex cancer[33] and identifying the functional gene sets that are highly associated with the clinical outcomes. However, the networks constructed with the knowledge-based interactions cannot reflect the heterogeneity of the cancer outcomes very well. Therefore, considering that the gene expression profiling measured by microarray or RNA-sequencing technology is one of the common elements for investigating the molecular mechanisms of cancers, it is essential to develop a strategy, which can construct the gene network only with the gene expression data instead of the known interactions in the database.

In our study, we presented a strategy to generate the potential connections among the genes in a specific cancer with the gene expression profiles, and adaptively divide the whole network into several functional modules, in which the genes directly connected with each other shared similar biological functions. Our proposed strategy was validated by using the microarray data of acute myeloid leukemia (AML), glioblastoma and neuroblastoma, as well as the RNA-sequencing data of glioblastoma (glioblastoma_seq). The gene set enrichment analysis for the genes in the functional modules clearly showed the similarities and differences among the three cancers. Kaplan-Meier analysis showed that the patients with diverse survival durations were well discriminated based on the expression profiles of the genes in functional modules, indicating that our strategy can successfully generate the networks with the gene expression data and extract the functional modules that are highly associated with cancer outcomes. Furthermore, the binary classification with the re-sampling method showed the robust performance of the predictive models on predicting the clinical outcomes of cancers.

Results

The gene networks constructed with the gene expression data displayed the scale-free characteristics. In our study, gene networks of three cancers were constructed by using graphical lasso based on gene expression profiles. All of them followed the power-law degree distribution, which were depicted in Supplementary Figs S1–S4. The networks were scale-free networks, in which most of the genes had a few neighborhoods, but some of them had many directly connected genes[34].

The top 20 genes ranked by degrees in the networks that were used for generating the seed modules. The top 20 hubs (genes) with high degrees were separately selected from the gene networks of AML, glioblastoma and neuroblastoma. The gene functions were searched in Tumorportal (http://www.tumorportal.org/) and GeneCards (http://www.genecards.org/) to obtain the gene functions.

Among the top 20 genes in AML gene network, four genes had been reported that were correlated with AML. The detailed information about the 20 genes was shown in Supplementary Table S1.

For the hubs in the glioblastoma gene network, ten genes had been reported to be associated with glioblastoma in previous researches. Details of these genes were listed in Supplementary Table S2.

As to the neuroblastoma gene network, the degrees of *NTRK1* and *CDH19* were much larger than those of other genes. Five out of 20 genes were directly related with neuroblastoma (Supplementary Table S3).

For the 20 hubs in the RNA-sequencing data of glioblastoma network, seven genes had reported to be related with glioblastoma (Supplementary Table S4).

Functional modules derived from the gene networks were associated with specific functions. In this study, two functional modules were separately generated for each gene network. We conducted Gene Ontology (GO) functional annotation and pathway analysis with the genes in each of the modules by using Database for Annotation, Visualization, and Integrated Discovery (DAVID) Functional Annotation Chart tool[35] and found the diverse functions between the modules of three cancers (Supplementary Tables S5–S12). For each gene network, the module related with the cancer progression was marked as module I, while another one was marked as module II.

The expression profiles of the genes in functional modules were correlated with the survival outcomes of patients. Hierarchical clustering analysis was conducted with the gene expression profiles as features to classify the patients. The 151 AML patients, 100 glioblastoma patients, 236 neuroblastoma patients, 138 RNA-sequencing data of glioblastoma patients, and the 237 independent testing set of neuroblastoma patients were separately grouped into two classes via hierarchical clustering (see Figs 1a,c, 2a,b, 3a,c, 4a,c and 5a,c).

Kaplan-Meier survival analysis was subsequently performed on the patients with AML, neuroblastoma, glioblastona_seq, and independent testing set of neuroblastoma (Figs 1b,d, 3b,d, 4b,d, 5b,d). The results suggested that the patients in AML, neuroblastoma, RNA-sequencing data of glioblastoma, and independent testing set of neuroblastoma datasets with shorter survival durations can be well discriminated by using the expression profiles of the genes in cancer-related functional modules.

Binary classification with the genes in functional modules. The averaged performance of the 100 times random re-sampling of the AML, glioblastoma, neuroblastoma, and the RNA-sequencing data of glioblastoma datasets was shown in Table 1. According to Table 1, all the features derived from module I achieved better performance than that derived from the module II, and the averaged ACCs for the module I of AML, glioblastoma, neuroblastoma, and RNA-sequencing data of glioblastoma were 0.919, 0.991, 0.899, and 0.902. The averaged MCCs for the functional modules of four datasets and the distribution for 100 times random re-sampling with the ACC and MCC were shown in Fig. 6. The performance of the independent testing set of neuroblastoma was shown in Table 2.

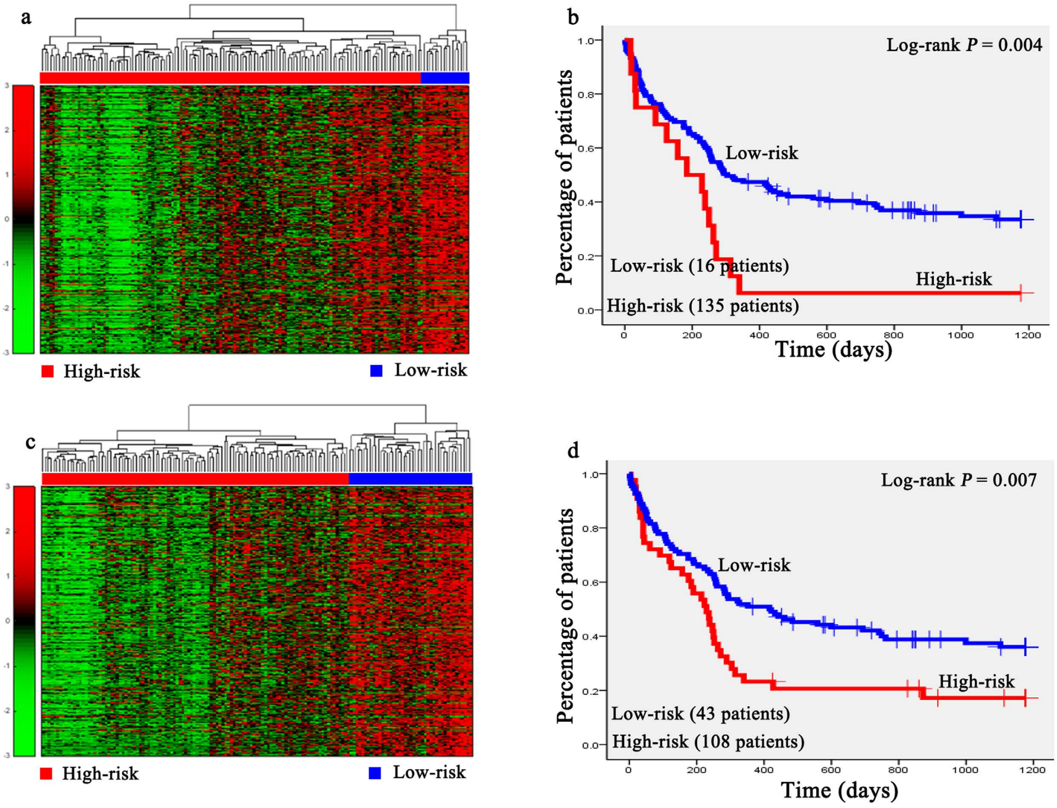

Figure 1. Unsupervised hierarchical clustering of genes in two modules and the Kaplan-Meier analysis of 151 patients with AML. (**a**,**c**) Unsupervised hierarchical clustering of tumor growth module and substance biosynthetic and metabolic module of AML. According to expression profiles of 1484 genes and 738 genes in two modules, 151 patients were clustered into two groups. The blue bar above the patients indicates the low risk group, the red bar indicates the patients with high risk. (**b**,**d**) Kaplan-Meier survival plots of low risk and high risk AML patients based on the hierarchical clustering results of (**a**,**b**) The gene expression signature of the patients in high risk (red lines) with a significantly increased risk of death when compared to low risk group (blue lines).

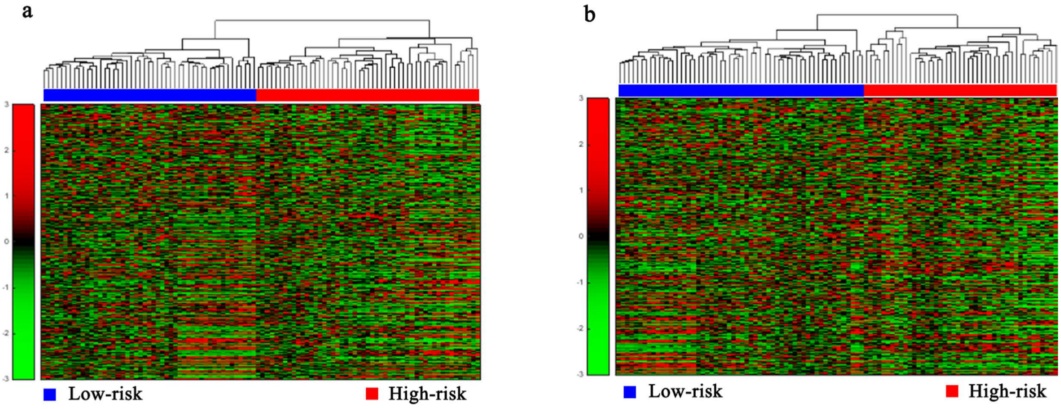

Figure 2. Unsupervised hierarchical clustering of genes in two modules of 100 patients with glioblastoma. (**a**,**b**) Unsupervised hierarchical clustering of the tumor proliferation and progression modules and cellular signaling transduction module of glioblastoma. According to expression profiles of 3710 genes and 1875 genes in two modules, 100 patients were clustered into two groups. The blue bar above the patients indicates the patients belong to the low risk group, the red bar indicates the patients with high risk.

Discussion

In this study, we proposed a strategy for gene network analysis, which is based on the gene expression data and can adaptively divide the whole network into several functional modules. The microarray data sets of three

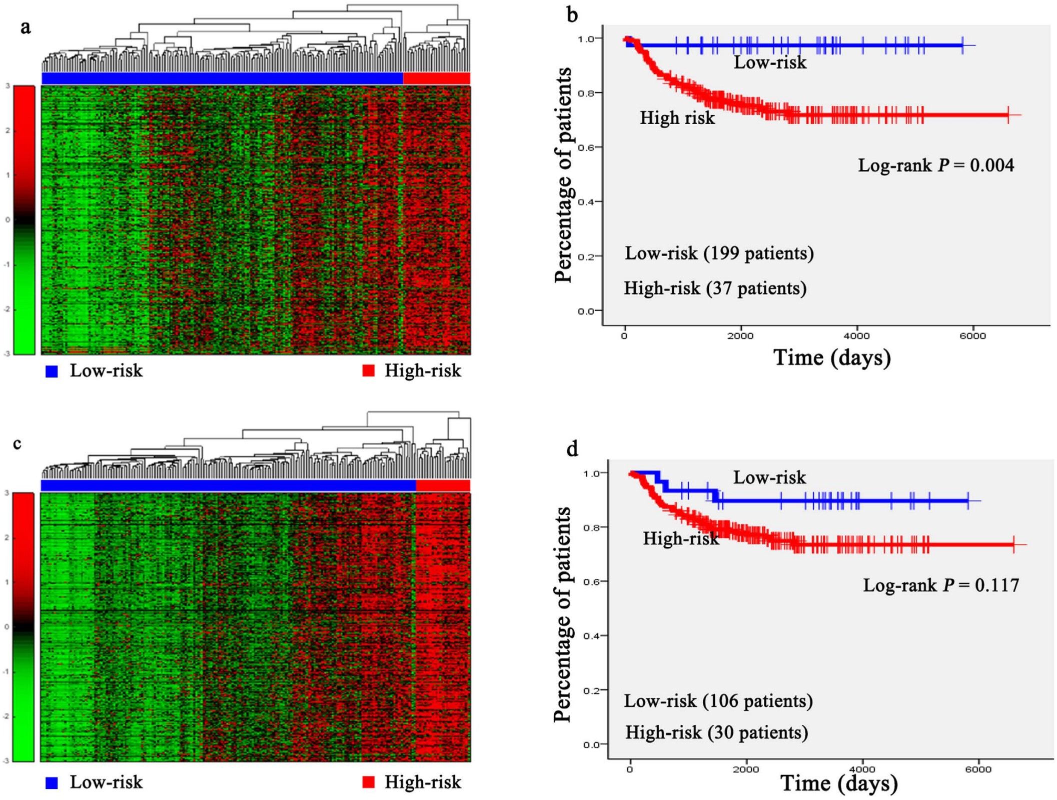

Figure 3. Unsupervised hierarchical clustering of genes in two modules and the Kaplan-Meier analysis of 236 patients with neuroblastoma. (**a**,**c**) Unsupervised hierarchical clustering of apoptosis and tumors migration module and nervous system development module of neuroblastoma. According to expression profiles of 3334 genes and 1739 genes in two modules, 236 patients were clustered into two groups. The blue bar above the patients indicates the low risk group, the red bar indicates the patients with high risk. (**b**,**d**) Kaplan-Meier survival plots of low risk and high risk neuroblastoma patients based on the hierarchical clustering results of (**a**,**b**) The gene expression signature in apoptosis and tumors migration module of the patients in high risk (red lines) with a significantly increased risk of death when compared to low risk group (blue lines).

cancers, namely AML, glioblastoma and neuroblastoma, as well as the RNA-sequencing dataset of glioblastoma, were applied to evaluate the performance of our strategy. Based on the gene expression profiles of the patients, we separately constructed the gene networks for four datasets and obtained two functional modules from each of the networks. In general, one functional module was related to the tumor initiation and progression (module I), and another one was mainly associated with the cellular process (module II). According to the hierarchical clustering and Kaplan-Meier analysis, the gene expression profiles in module I achieved better discrimination of the patients' risk group. To tested whether the clustering results achieved with the genes in functional modules are better than those obtained with the randomly selected genes by using the Monte Carlo sampling procedure. In total, 1484, 3710 and 3334 genes were grouped into the cancer related modules for the datasets of AML, glioblastoma and neuroblastoma, respectively, and used in the subsequent clustering analysis and survival analysis. So, for each of the datasets, we randomly selected the same number of genes from the gene network and used them for clustering analysis and survival analysis. This procedure has been repeated for 100 times and the p values obtained by log-rank test (conducted in survival analysis) were shown in Supplementary Fig. S5. From the figure, we find that the p values ($p = 0.004$, 0.004, and 0.001 for AML, neuroblastoma, and glioblastoma, resp.) obtained by using the genes in module I were less than those (average p value = 0.057, 0.125, and 0.160 for AML, neuroblastoma, and glioblastoma, resp.) obtained in the 100 re-sampling procedures. We also used all the differentially expressed genes (DEGs) for the clustering analysis and the survival analysis, the log-rank p values were 0.004, 0.008, and 0.211 for AML, neuroblastoma, and glioblastoma, respectively. We considered that, to some extent, the genes in the functional modules performed better than those selected by chance in clustering the patients into high/low risk groups.

The gene networks were constructed by using graphical lasso based on gene expression profiles. In the AML gene networks, the degrees of a few genes were very larger than the most genes in the network, which were shown in Fig. 7. The node degree distribution of the AML gene network followed a power-law distribution (Supplementary Fig. S1), which was a characteristic of scale-free networks[36]. The other three gene networks also showed the same property (Supplementary Figs S2–S4). The top 20 genes ranked by degrees and their first neighbors in the networks were the basis of our strategy to identify the functional modules. Thus, the scale-free topology of the gene networks was necessary in this study.

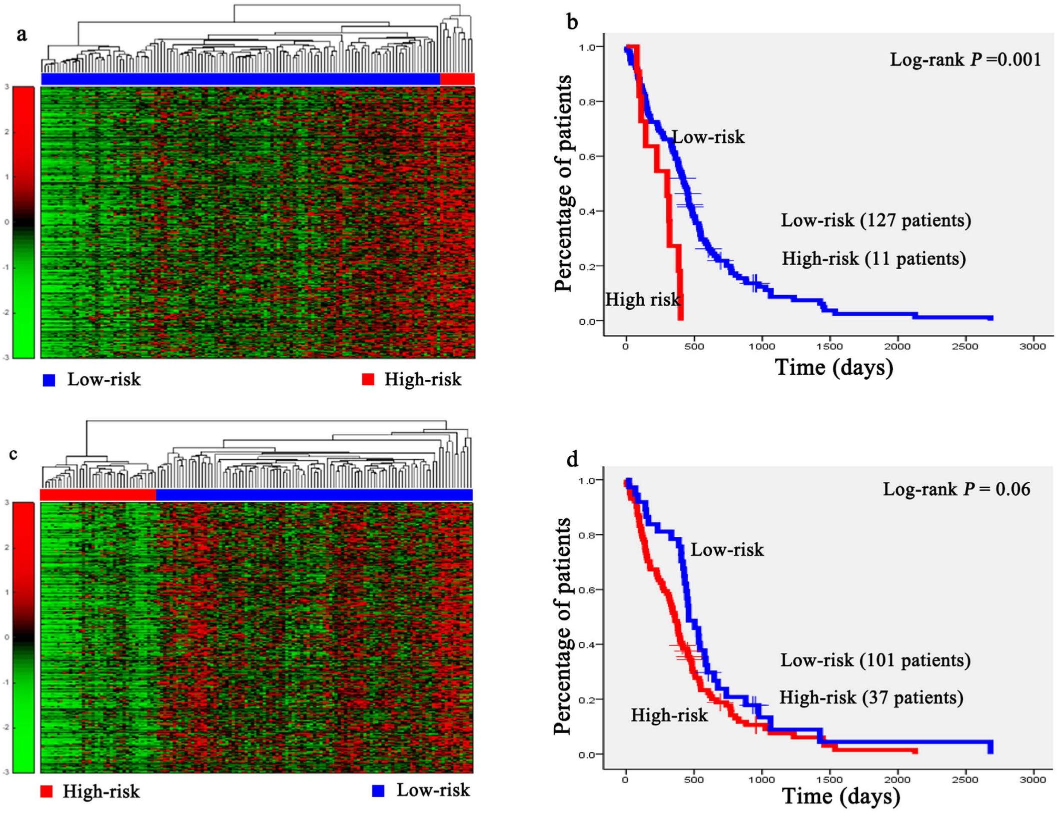

Figure 4. Unsupervised hierarchical clustering of genes in two modules and the Kaplan-Meier analysis of 138 glioblastoma patients of the RNA-sequencing dataset. (**a**,**c**) Unsupervised hierarchical clustering of apoptosis and cell death regulation module and substance biosynthesis and metabolism module of RNA-sequencing data of glioblastoma. According to expression profiles of 1417 genes and 794 genes in two modules, 138 patients were clustered into two groups. The blue bar above the patients indicates the low risk group, the red bar indicates the patients with high risk. (**b**,**d**) Kaplan-Meier survival plots of low risk and high risk RNA-sequencing data of glioblastoma patients based on the hierarchical clustering results of (**a**,**b**). The gene expression signature in apoptosis and cell death regulation module of the patients in high risk (red lines) with a significantly increased risk of death when compared to low risk group (blue lines).

In the gene network of AML, *SOCS2* and *CDK6* were involved in two modules. The overexpression of *SOCS2* in AML patients indicates the favorable prognosis[37]. *CDK6* plays an vital role in mixed-lineage leukemia fusions in myeloid leukemogenesis, which is associated with poor prognosis of patients[38]. In the module related to the tumor growth (Supplementary Table S5), the GO terms of damaged DNA binding was significantly enriched ($p = 2.09 \times 10^{-2}$), which promoted to tumor suppression[39]. The enriched biological process (BP) of glycolysis ($p = 4.7 \times 10^{-2}$) is usually active in tumor cells and the intermediates were involved in tumor tissues[40]. The oxidative phosphorylation pathway ($p = 1.59 \times 10^{-9}$) was significantly enriched in tumor growth module, which decreased the availability of ATP that associated with malignancies and tumor cell expansion and was reported to be one of the key progress in tumor growth[41]. In particular, the level of oxidative phosphorylation pathway might leads to the changes of mitochondrial functions, which was the character of the hematological cell malignancies, AML was included[42]. In the module related to the biosynthesis and metabolism of substance, the pathways of steroid biosynthesis ($p = 3.50 \times 10^{-3}$), pyruvate metabolism ($p = 1.81 \times 10^{-2}$), and amino sugar and nucleotide sugar metabolism ($p = 2.64 \times 10^{-2}$) were significantly enriched, which mainly contributed to the biosynthetic and metabolic processes of the proteins, steroids and pyruvate. The GO terms and pathways enriched with the genes in this module were listed in Supplementary Table S6. From the results of Kaplan-Meier survival analysis (Fig. 1b,d), we found that the difference of survival durations of the patients between high- and low-risk groups separated by the gene expression profiles in the module related to tumor growth was more significant than that separated by the gene expression profiles in another module, indicating that the genes involved in the initiation and progression of AML were more suitable for the prognosis of patients.

For glioblastoma, in the module of tumor proliferation and progression, the hub *FSTL5*, as a candidate gene for tumor-suppressor, encodes a secretory glycoprotein and its expression level is highly correlated with tumor size. *IL13RA2* is a tumor antigen-like factor, which is highly expressed in the glioblastoma patients[43]. Two enriched GO terms named cell morphogenesis involved in differentiation ($p = 1.24 \times 10^{-5}$) and cell migration ($p = 2.59 \times 10^{-6}$) mainly contribute to the tumor progression and migration. MAPK signaling pathway ($p = 4.31 \times 10^{-4}$) and the pathway in cancer ($p = 3.48 \times 10^{-3}$) were also significantly enriched in this module (Supplementary Table S7).

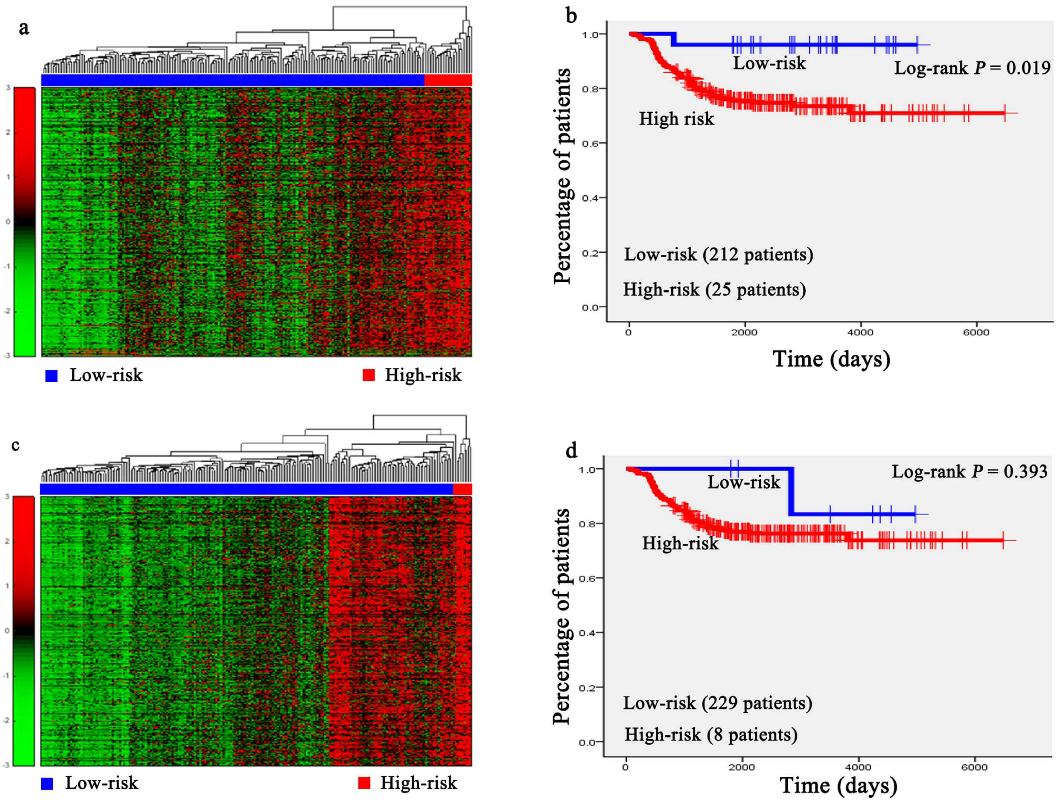

Figure 5. Unsupervised hierarchical clustering of genes in two modules and the Kaplan-Meier analysis of the independent testing set of 237 neuroblastoma patients. (**a**,**c**) Unsupervised hierarchical clustering of two modules of independent testing set. According to expression profiles of 3334 genes and 1739 genes in two modules, 237 patients were clustered into two groups. The blue bar above the patients indicates the low risk group, the red bar indicates the patients with high risk. (**b**,**d**) Kaplan-Meier survival plots of low risk and high risk neuroblastoma patients based on the hierarchical clustering results of (**a**,**b**). The gene expression signature in module I of the patients in high risk (red lines) with a significantly increased risk of death when compared to low risk group (blue lines).

	AML		Glioblastoma		Neuroblastoma		Glioblastoma_seq	
	Module I	Module II	Module I	Module II	Module I	Module II	Module I	Module II
SPC	0.925	0.878	0.987	0.967	0.885	0.922	0.888	0.882
SEN	0.912	0.855	0.996	0.999	0.913	0.850	0.915	0.889
ACC	0.919	0.866	0.991	0.983	0.899	0.886	0.902	0.885
MCC	0.841	0.741	0.984	0.968	0.813	0.785	0.810	0.777

Table 1. The averaged performance of the 100 times random re-sampling with the SVM models.

The enriched Wnt signaling pathway ($p = 4.73 \times 10^{-5}$) had been reported in previous study that it might contribute to glioma cell proliferation and impaired apoptosis, which allowed for the progression to glioblastoma[44]. Genes in axon guidance pathway ($p = 3.39 \times 10^{-3}$) were reported to be differentially expressed in the patients with Parkinson disease compared to the healthy ones, which are the signs of the occurrence of the brain disorder[45]. In the module about cellular signaling transduction, the enriched GO terms named intracellular signaling cascade ($p = 3.60 \times 10^{-7}$), regulation of cell communication ($p = 1.12 \times 10^{-5}$) and ion transport ($p = 1.30 \times 10^{-5}$) contribute to the transduction among the cells. The most significantly enriched calcium signaling pathway ($p = 1.82 \times 10^{-4}$) in this module, was involved in regulating neural functions such as brain rhythms generation and information transduction in the processes of memory and learning[46]. The pathways about cellular signaling transduction were also significantly enriched with the genes in this module (Supplementary Table S8).

In the gene network of neuroblastoma, one functional module was about cell apoptosis and tumor migration and another was about the development of nervous system. The hub *NTRK1* was involved in two functional modules. It is high expressed in the low-risk group of neuroblastoma patients, indicating the favorable prognosis of neuroblastoma patients[47]. In the module about cell apoptosis and tumor migration, Pathway of apoptosis was a genetically controlled mechanism of cell death involved in the regulation of tissue homeostasis. MAPK signaling pathway ($p = 4.23 \times 10^{-3}$) was activated in human cancers and resulted in malignant phenotypes[48]. The pathway of cell adhesion molecules (CAMs) ($p = 2.27 \times 10^{-8}$) was involved in tumors migration[49]. All the enriched GO

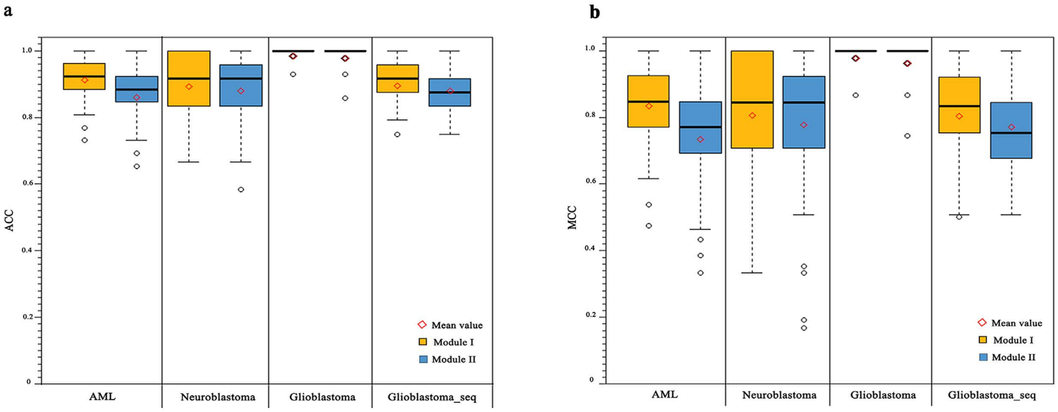

Figure 6. Performance of the 100 times re-sampling for each module of four datasets. (**a**) The distribution of ACC values of the 100 times re-sampling for module I (left, yellow) and module II (right, blue) of four datasets. (**b**) The distribution of MCC values of the 100 times re-sampling for module I (left, yellow) and module II (right, blue) of four datasets. ◇ is the mean value of the 100 times random re-sampling of each module.

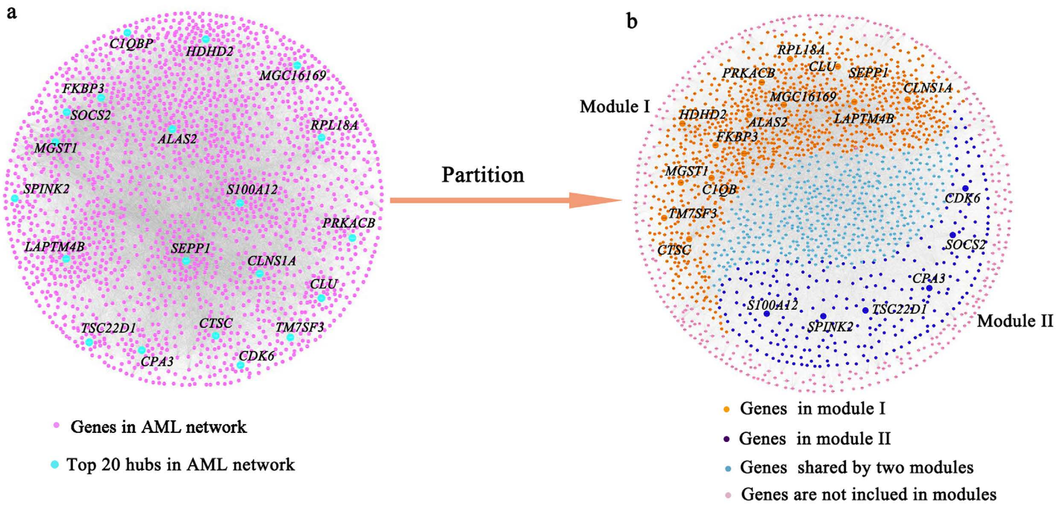

Figure 7. Network and the functional modules of AML. (**a**) The visualization of AML gene network. (**b**) The visualization of the functional modules in the gene network.

Neuroblastoma	SPC	SEN	ACC	MCC
Module I	0.910	0.500	0.865	0.375
Module II	0.886	0.500	0.849	0.333

Table 2. Performance of the independent testing set of the 237 neuroblastoma patients.

terms and pathways for this module were listed in Supplementary Table S9. In the module about the development of nervous system, the enriched GO terms of the ectoderm development ($p = 8.88 \times 10^{-6}$) and organ development ($p = 1.26 \times 10^{-8}$) are the final steps in the development of nervous system. The pathway of neuroactive ligand-receptor interaction ($p = 1.11 \times 10^{-5}$) was associated with the neuronal function[50]. The pathway of cytokine–cytokine interactions ($p = 1.08 \times 10^{-3}$) is crucial during immunological and inflammatory responses in disease[51]. The detailed results of the DAVID analysis for this module were listed in Supplementary Table S10. Based on the expression profiles of the genes in these two modules, the patients were significantly divided into high-risk and low-risk groups by hierarchical clustering analysis (Fig. 3b,d). Note that the difference of survival durations of the patients between high- and low-risk groups separated by the gene expression profiles in the module related to cell apoptosis and tumor migration was more significant than that separated by the gene expression profiles in another module.

For the gene network constructed based on RNA-sequencing data of glioblastoma, one of the functional modules was mainly about apoptosis and programmed cell death regulation, and the other one was about substance biosynthesis and metabolism. Expression of *PI3* could be induced by inflammatory mediators such as tumor necrosisfactor and the interleukin 1 beta[52]. The *L1CAM* could stimulate glioblastoma cells motility and proliferation[53]. In the module about apoptosis and programmed cell death regulation, GO terms enriched in it were tumor necrosis factor (TNF) binding ($p = 2.71 \times 10^{-2}$), tumor necrosis factor receptor superfamily binding ($p = 2.90 \times 10^{-2}$). TNF has been reported involved in the cancers development and progression in preclinical models[54]. The pathways enriched in this modules were cytokine–cytokine interactions ($p = 4.06 \times 10^{-12}$), apoptosis ($p = 1.34 \times 10^{-3}$) (Supplementary Table S11). In the module of substance biosynthesis and metabolism, the enriched GO terms were sulfuric ester hydrolase activity ($p = 6.83 \times 10^{-5}$), cytokine activity ($p = 2.07 \times 10^{-4}$), the enriched pathway were ECM-receptor interaction($p = 3.86 \times 10^{-5}$), keratan sulfate biosynthesis ($p = 3.59 \times 10^{-3}$), and fructose and mannose metabolism ($p = 2.20 \times 10^{-2}$). The results of DAVID analysis were shown in Supplementary Table S12. Fig. 4b,d showed that the difference of survival durations of the patients between high- and low-risk groups separated by the gene expression profiles in the module related to apoptosis and programmed cell death regulation was more significant than that separated by the gene expression profiles in another module, indicating that the genes involved in the cancers progression were more suitable for the prognosis of patients. Our strategy could be applied to both microarray and RNA-sequencing data.

We further investigated the differences and similarities of the modules among three microarray datasets. The overlaps of the genes in the modules that related to the cancer initiation and progression and cellular process were shown in Supplementary Fig. S6. It can be seen that 789 genes (21.3%) in glioblastoma overlapped with neuroblastoma in cancer initiation and progression (Supplementary Fig. S6a), while 353 genes (9.5%) in glioblastoma overlapped with AML. Likewise, 109 genes (5.8%) in glioblastoma overlapped with neuroblastoma in cellular process (Supplementary Fig. S6b), while 79 genes (4.2%) in glioblastoma overlapped with AML. It indicates that the molecular basis of glioblastoma was similar to that of neuroblastoma.

Furthermore, we evaluated the genes predictive ability for the patients' cohort in the functional modules by using SVM modeling. To avoid the overfitting, feature selection was applied to reduce dimension of the genes for each dataset. As the models are often not robust, re-sampling is a good manner to test the robustness of the models, and Li *et al.* achieved the high quality cancer signature genes identification by MSS algorithm with re-sampling test[55,56]. According to averaged values of the evaluation parameters in Table 1, the features selected from functional modules achieved a good performance for predicting the patients risk group. The ACC and MCC distribution of 100 times re-sampling (Fig. 6) suggested that our models were robust to predict the patients' cohorts.

In summary, our strategy can generate the gene network based on the gene expression data and adaptively divide the whole network into the functional modules. To facilitate the clinical application, the gene network can be constructed only with the gene expression data by using our proposed strategy. Considering the limited information of gene expression data, not all the genes in the functional modules are necessarily associated with cancers. Therefore, before using these genes for the subsequent analysis and prediction, it is still necessary to refine the genes by using the knowledge-based datasets, such as the curated human signaling network (http://www.cancer-systemsbiology.org/dataandsoftware.htm), which is created by manually collecting the data sources from multiple pathway-related databases (signaling network of BioCarta, CST Signaling pathways, Pathway Interaction, iHOP) and literatures, and successfully applied in the identification of cancer biomarkers[55,57–59]. In this study, the two modules extracted from the gene networks of three cancers exhibited distinct functions, which were associated with the cancer initiation and progression, and cellular process, respectively. The gene expression profiles in both two modules were significantly related with the pathogenesis and prognosis of these cancers, and can well discriminate the patients with shorter survival durations from the cancer patient cohort.

Material and Methods

Datasets. The microarray datasets of AML, glioblastoma and neuroblastoma used in this study were available at the Gene Expression Omnibus (GEO) database and the accession numbers were GSE12417, GSE4271 and GSE49710, respectively. The RNA-sequencing data was downloaded from the Cancer Genome Atlas (TCGA). Neuroblastoma dataset contains 473 samples that are independently profiled with Agilent-020382 Human Custom Microarray 44 k. AML dataset is consist of 151 samples and profiled with Affymetrix microarray HG-U133A&B and HG-133Plus2. Glioblastoma dataset consists of 100 samples and profiled with Affymetrix microarray HG-U133A. The RNA-sequencing data of glioblastoma was generated by Illumina HiSeq RNASeqV2. The raw count for each of genes (data in level 3) was used as expression value. The RNA-sequencing data of glioblastoma dataset consisted of 138 samples. In addition, for the 473 neuroblastoma patients, 236 samples are used to construct gene network, and the others are used as the independent testing set.

For the multiple probesets that mapped to the same HUGO gene symbol, we only kept the probeset with maximum value as the expression level of this gene. The patients in the datasets of AML, neuroblastoma, and the RNA-sequencing data of glioblastoma were divided into two groups according to the survival days. The patients of AML with survival days longer than one year were assigned to the low risk group. Otherwise, patients were assigned to the high risk group. For the neuroblastoma and RNA-sequencing data of glioblastoma datasets, the patients with survival days longer than two years were assigned to the low risk group and the patients with survival days less than two years were assigned to the high risk group. If a patient with the survival days were less than the cut-off days and he/she was still alive, he/she would be removed from the dataset. For the glioblastoma patients, the patients belong to ProNeural (PN) subcategory were assigned to the low risk group and the rest were assigned to the high risk group[60].

After the preprocess of the raw four data sets, 14,894, 14,894, 19,860, and 20,531 genes were included in AML, glioblastoma neuroblastoma, and RNA-sequencing data of glioblastoma datasets, too many genes made the construction of gene networks by graphical lasso impossible, thus we need to select a part of genes for following study. Student's t-test was used to identify the DEGs and those with p value < 0.05, and the false discovery rate was not

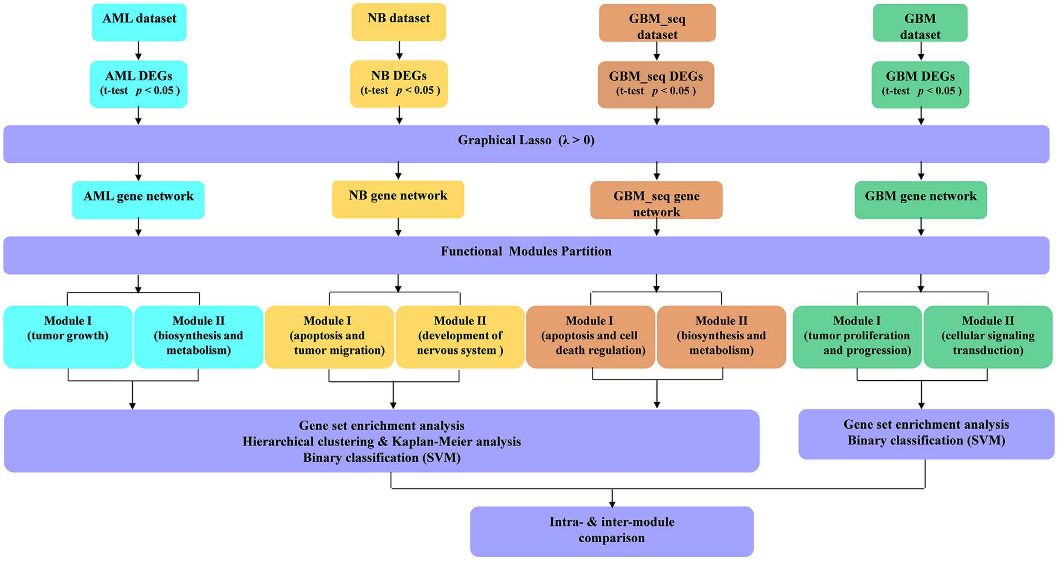
AML dataset
NB dataset
GBM_seq dataset
GBM dataset
AML DEGs (t-test p < 0.05)
NB DEGs (t-test p < 0.05)
GBM_seq DEGs (t-test p < 0.05)
GBM DEGs (t-test p < 0.05)
Graphical Lasso (λ > 0)
AML gene network
NB gene network
GBM_seq gene network
GBM gene network
Functional Modules Partition
Module I (tumor growth)
Module II (biosynthesis and metabolism)
Module I (apoptosis and tumor migration)
Module II (development of nervous system)
Module I (apoptosis and cell death regulation)
Module II (biosynthesis and metabolism)
Module I (tumor proliferation and progression)
Module II (cellular signaling transduction)
Gene set enrichment analysis
Hierarchical clustering & Kaplan-Meier analysis
Binary classification (SVM)
Gene set enrichment analysis
Binary classification (SVM)
Intra- & inter-module comparison

Figure 8. A schematic view of our work in this article as part of the whole research project for functional modules partition. AML: acute myelocytic leukemia. GBM: glioblastoma. NB: neuroblastoma. GBM_seq: RNA-sequencing data set of the glioblastoma. DEGs: differentially expressed genes. SVM: support vector machine.

controlled. As a result, 10,184, 2966, 7373 and 2528 DEGs were selected for neuroblastoma, AML, glioblastoma and RNA-sequencing data of glioblastoma, respectively. The overview of this work was shown in Fig. 8.

Gene-gene network construction. In this work, the most based strategy, i.e. using the inverse of the covariance matrix, was used to build the network. However, the number of DEGs for four datasets are all larger than the number of samples, thus we got a noninvertible covariance matrix[61]. To address this problem, we applied the graphical lasso, an algorithm of estimation of a sparse inverse covariance matrix using L_1 penalty, to learn the structure in an undirected graphical model[62]. Here, we definite P and N as the number of the gene and sample respectively, $\sum$ is a covariance matrix which size is P×N.

Let $\theta = \sum^{-1}$, and let S be the empirical covariance matrix, the problem is to maximize the penalized log-likehood

$$\log \det\theta - \mathrm{tr}\,(S\theta) - \rho||\theta||_1 \tag{1}$$

Here tr represents the trace and ρ is a nonnegative tuning parameter; $||\theta||_1$ is the L_1 norm, the sum of the absolute values of the element θ.

The sparse of the θ is decided by L_1 norm, when ρ is sufficiently large, the estimation of the inverse covariance matrix would be sparse enough, which can help us to find the significant strong interaction among the elements input. While the small of the ρ value means the covariance matrix that we estimated would better explain the model we construct from the data. The topology of the gene networks was estimated by setting the edges to correspond to the nonzero elements of the inverse covariance matrix θ[63].

In this study, for each dataset, the penalty values for three datasets should meet the needs of modules partition. The penalty parameters were ranking from 0 to 1.0, and the step size was 0.1. The sparsity of the network was proportional to the penalty parameter, the time of computation was increasing with the sparsity of the network as well as the number of elements of the datasets. The penalty parameters for AML, glioblastoma, neuroblastoma and the RNA-sequencing data of glioblastoma were 0.2, 0.7, 1.0 and 0.6, respectively.

Functional module partition based on the network. For a constructed network (also called graph) G = (V, E), the vertexes (V) and the edges (E) are defined previously, i.e. genes and their relationships. Firstly, we calculated the degree of all the genes and ranked them, the first 20 genes with the largest degree are picked out as the hubs $H = \{h\}_{i=1,\ldots,n} \in V$. Then the modules $U = \{U_i\}_{i=1,\ldots n}$ could be initialized based on the hub genes:

$$U_i = \{\text{The nodes in V adjacent with } h_i\} \bigcup h_i \tag{2}$$

That is, the genes which adjacent with a hub gene are in an initialized module.

If the intersection of two (or more) initialized modules is null, we defined the two (or more) modules as a seed module set. Therefore, the seed modules $S = \{S_i\}_{i=1,\ldots,m}$ could be defined as

$$S_i = \{s_{ik},\ k = 1,\ \ldots,\ \alpha|\ \text{for}\ \forall j,\ k,\ s_{ik} \in U,\ s_{ij} \bigcap s_{ik} = \varnothing\} \tag{3}$$

For convenience, we focused on the first seed module set S_1, the rest modules could be represented as R_1, i.e.

$$R_i = \left\{r_{ij},\ j = 1,\ \ldots,\ \beta|\ r_{ik} \in C_U S_i\right\} \tag{4}$$

Then we could merge the elements in R_1 into S_1 by calculating the overlaps between each r_{ij} and s_{ik}:

$$\text{overlap} = \frac{\left|s_{ik} \cap r_{ij}\right|}{\left|s_{ik} \cup r_{ij}\right|} \tag{5}$$

The symbol '||' means the card of a set, i.e. the number of the elements in a set. Every r_{ij} could be merged into a s_{ik} with the largest overlap. Note that after each merging, the merged s_{ik} was updated as the merged set.

After merging, R_1 was merged into S_1, the merged set is represented as C_1. Repeat the steps, all the elements in S could be updated as $C = \{C_i\}$. The elements in C_i could be regard as a candidate of the partitioned modules. To find the best partition modules, an evaluation method is needed. In this work, we still evaluated $C_i = \{c_{i1}, \ldots, c_i\alpha\}$ by calculating the average overlap:

$$E_i = \frac{\sum_{m,n}\frac{|c_{im} \cap c_{in}|}{\min(|c_{im}|, |c_{in}|)}}{\binom{\alpha}{2}} m \neq n,\ m = 1,\ \ldots,\ \alpha \text{ and } n = 1,\ \ldots,\ \alpha \tag{6}$$

$\binom{\alpha}{2}$ is the combination number, namely, Binomial coefficient. In our opinion, if two modules are well divided, the average overlap of the modules will be small. Thus, C_i with the smallest E_i is selected as the best partitioned modules. Moreover, the smallest E_i might be not unique (for instance, two C_i have the same E_i), thus the methods might get more than one result based on different seed module sets. However, in this work, even we encounter this situation, that is, two different seed module sets, S_i and S_j, got the same smallest evaluation number , but the updated module sets, C_i and C_j, are completely the same.

Parameter Optimization of the penalty parameter.

Note that seed module set ($|S| \geq 1$) is indispensable, the constructed network would contain at least two initialized modules which have no any common genes. Therefore, the parameter optimization is necessary. Since the limitation of GO analysis, a module which contains more genes could get the enrichment result. Thus our strategy in this work is to get the parameter which could reserve more edge and ensure the constructed network contain at least one seed module set. In this study, the penalty parameters were ranking from 0 to 1.0, and the step size was 0.1. The sparsity of the network was proportional to the penalty parameter, the time of computation was increasing with the sparsity of the network as well as the number of elements of the datasets. The penalty parameters for AML, glioblastoma, neuroblastoma and the RNA-sequencing data of glioblastoma were 0.2, 0.7, 1.0 and 0.6, respectively.

The pseudocodes of section 2 to section 4 are listed in Table 3. Moreover, to provide easy use of this method, the web serve of functional modules partition strategy is available at http://scu-cheminfo.com/GFMP/. The MATLAB code of the strategy and the datasets are also included in this website.

GO Term Enrichment and Pathway Enrichment Analysis.

The Gene Ontology (GO) contains three categories, namely Biological process (BP), Molecular function (MF) and Cellular component (CC), which can consistently describe and annotate gene products[64]. In our study, the enrichment of GO terms was performed by using DAVID (The database for Annotation, Visualization and Integrated Discovery) Functional Annotation Chart tool[35,65] (http://david.abcc.ncifcrf.gov/). The GO terms are organized as a tree structure, wherein term specificity increases and genome coverage decreases as one moves down the tree structure. Level 1 provides the highest list coverage with the least amount of term specificity and level 5 provides the least amount of coverage with the highest term specificity[35,66]. Here, in order to avoid very general and uninformative GO terms, we only used the GO terms in level 4. The *p* values were corrected for multiple testing using the Bonferroni procedure and transformed by taking the -log10 for easier visualization[66].

Based on the DEGs, the pathway enrichment analysis calculates the hypergeometric distribution between the DEGs and pathway and gives a *p* value for each pathway that contains any DEGs. The small *p* value suggests that the DEGs are enriched in the pathway. As DAVID consists of an integrated biological knowledge base and analytic tools aimed at systematically extracting biological meaning from large gene lists, we directly used it to identify the over-represented KEGG pathway[67]. To guarantee the statistically significant in this study, we kept the enriched GO terms and pathways with the *p* value less than or equal to 0.05.

Feature selection for binary classification.

As the number of genes in each functional module is larger than the number of samples in four datasets, overfitting is a challenge when all genes in modules are used for modeling. Thus, the feature selection methods are applied before building modules. 30 genes are selected from every functional module. Then grid search is used for parameter optimization, and the built modules are determined by 5-fold cross-validation. The whole process of the feature selection and modeling are achieved by using a Parallel Machine Learning toolbox for data classification and regression (PML) platform (http://cic.scu.edu.cn/pml/)[68]. Among the thirteen features used selection methods, SVMAttributeEval (An evaluater which select the feature by using an SVM classifier) outperforms other methods. The recursive feature elimination (RFE) evaluates the features by using SVM classifier. Features are assigned a parameter by using the algorithm SVM-train, and then the parameter is used to calculate the weight of the feature by a decision function. RFE is the ranking criterion of the weight of feature. Feature selection is according to the ranks of the features based on the FRE[69].

Input:	
G	*Network*
$\boldsymbol{H}=\{h_i\}$ *where* $i=1, \ldots, n$	*The hub genes*
Output:	
$\boldsymbol{B}=\{\boldsymbol{B}_i\}$	*The best module set(s)*
Intermediate variable:	
$\boldsymbol{U}=\{\boldsymbol{U}_i\}$ *where* $i=1, \ldots, n$	*All of the modules*
$\boldsymbol{S}=\{\boldsymbol{S}_i\}$ *where* $i=1, \ldots, m$	*Seed modules sets*
$\boldsymbol{R}=\{\boldsymbol{R}_i\}$ *where* $i=1, \ldots, \beta$	*The supplementary set of Seed modules sets*
$\boldsymbol{C}=\{\boldsymbol{C}_i\}$ *where* $i=1, \ldots, m$	*The combined modules*
$\boldsymbol{E}=\{E_i\}$ *where* $i=1, \ldots, m$	*The evalution values of the combined modules*

```
Find all the $\boldsymbol{U}_i = \{\text{The nodes adjacent with } h_i\} \cup h_i$, where $i = 1, \ldots, n$
Find all the $\boldsymbol{S}_i = \{\boldsymbol{s}_{ik}, k = 1, \ldots, \alpha \mid \text{for } \forall\, j, k,\ \boldsymbol{s}_{ik} \in \boldsymbol{U}, \boldsymbol{s}_{ij} \cap \boldsymbol{s}_{ik} = \varnothing\}$
  where $i=1, \ldots, m$
For i from 1 to m
    $\boldsymbol{R}_i = \{\boldsymbol{r}_{ij}, j = 1, \ldots, \beta \mid \boldsymbol{r}_{ik} \in C_{\boldsymbol{U}}\boldsymbol{S}_i\}$
    For j from 1 to $\beta$
        For k from 1 to $\alpha$
            $o_k = \frac{|s_{ik} \cap r_{ij}|}{|s_{ik} \cup r_{ij}|}$
    end loop k
    $k_{Max} = \inf_k\{\mathrm{o}_\mathrm{k} = \max(o_1, \ldots, o_\alpha)\}$
    $\mathbf{S}_{\mathrm{ik}_\mathrm{Max}} = \mathbf{S}_{\mathrm{ik}_\mathrm{Max}} \cup \mathbf{r}_\mathrm{ij}$
    Remove $\mathbf{r}_\mathrm{ij}$ in $\boldsymbol{R}_i$
  end loop j
  $\boldsymbol{C}_\mathrm{i}=\boldsymbol{S}_i$
end loop i
For i from 1 to m
  Get combined modules $\mathbf{C}_\mathrm{i}=\{\mathbf{c}_{\mathrm{i}1}, \ldots, \mathbf{c}_{\mathrm{i}\alpha}\}$
  $\mathrm{E}_\mathrm{i} = \frac{\sum_{m,n}\frac{|\boldsymbol{c}_{im} \cap \boldsymbol{c}_{in}|}{\min(|\boldsymbol{c}_{im}|, |\boldsymbol{c}_{in}|)}}{\binom{\alpha}{2}}\ m \neq n, m = 1, \ldots, \alpha \text{ and } n = 1, \ldots, \alpha$
end loop i
$\mathrm{i}_\mathrm{best} = \inf_i\{\mathrm{E}_\mathrm{i} = \min(\mathrm{E}_1, \ldots, \mathrm{E}_m)\}$
$\mathbf{B} = \{\mathrm{C}_{\mathrm{i}_\mathrm{best}}\}$
```

Table 3. The pseudocode of functional modules partition scheme. Note: The i_best might not unique, Thus the $|\mathbf{B}|$ could be larger than 1.

Machine learning with re-sampling procedure.

To evaluate the robustness, the re-sampling mechanism is used in this work. SVM is a supervised learning method and often used for binary classification[70]. The random re-sampling was repeated 100 times to select training set and valid set from the AML, neuroblastoma, and RNA-sequencing data of glioblastoma datasets, and the averaged values were used for estimating the robust of the models. The LIBSVM in MATLAB was used for modeling[71]. The training models were determined by grid search and 5-fold cross-validation. The ratio of the positive samples and negative samples was 1:1. The ratio of the training set and validation set was 4:1. Here, sensitivity (SEN), specificity (SPC), accuracy (ACC), and Mathew's correlation coefficient (MCC) were used as evaluation parameters. The related formulas are as follow:

$$SEN = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{7}$$

$$SPC = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{8}$$

$$ACC = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{9}$$

$$MCC = \frac{(\mathrm{TP} \times \mathrm{TN}) - (\mathrm{FN} \times \mathrm{FP})}{\sqrt{(\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP}) \times (\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FN})}} \tag{10}$$

where *TP* (True Positive) is the number of the correctly predicted positive sample, *TN* (True Negative) is the number of the correctly predicted negative sample, *FP* (False Positive) is the number of the incorrectly predicted positive sample, and *FN* (False Negative) is the number of the incorrectly predicted negative sample.

Hierarchical Clustering survival analysis and network visualization. For the hierarchical clustering analysis, only genes expression profiles in modules were used for clustering. The whole procedure was conducted in MATLAB 8.0. The Kaplan-Meier survival analysis was conducted by using IBM SPSS Statistics 19. The network visualization was conducted by Cytoscape 3.2.0.

Data availability. The web serve of functional modules partition strategy is available at http://scu-cheminfo.com/GFMP/. The MATLAB code of the strategy and the datasets are also included in this website. We also upload the codes and the data to GitHub (https://github.com/limlcic/GFMP).

References

1. Walker, B. A. *et al.* APOBEC family mutational signatures are associated with poor prognosis translocations in multiple myeloma. *Nat. Commun.* **6,** 6997 (2015).
2. Thunnissen, E., Oord, K. V. D. & Bakker, M. D. Prognostic and predictive biomarkers in lung cancer. A review. *Archiv Für Pathologische Anatomie Und Physiologie Und Für Klinische Medicin* **464,** 347–358 (2014).
3. Kosmider, O. *et al.* Mutation of the colony-stimulating factor-3 receptor gene is a rare event with poor prognosis in chronic myelomonocytic leukemia. *Leukemia* **27,** 1946–1949 (2013).
4. Sato, T. *et al.* PRC2 overexpression and PRC2-target gene repression relating to poorer prognosis in small cell lung cancer. *Sci. Rep.* **3,** 1911 (2013).
5. Arteaga, C. L. *et al.* Treatment of HER2-positive breast cancer: current status and future perspectives. *Nat. Rev. Clin. Oncol.* **9,** 16–32 (2011).
6. Shi, L. *et al.* The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat. Biotechnol.* **28,** 827–838 (2010).
7. Badertscher, M., Bischofberger, K., Munk, M. E. & Pretsch, E. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J. Natl. Cancer Inst.* **98,** 1183–1192 (2006).
8. Croce, C. M. Oncogenes and cancer. *N. Engl. J. Med.* **358,** 502–511 (2008).
9. Yang, Y. *et al.* Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nat. commun.* **5,** 3231 (2014).
10. Aderem, A. Systems biology: its practice and challenges. *Cell* **121,** 511–513 (2005).
11. Ge, H., Walhout, A. J. & Vidal, M. Integrating 'omic'information: a bridge between genomics and systems biology. *Trends Genet.* **19,** 551–560 (2003).
12. Albert, R. Scale-free networks in cell biology. *J. Cell Sci.* **118,** 4947–4957 (2005).
13. Furlong, L. I. Human diseases through the lens of network biology. *Trends Genet.* **29,** 150–159 (2013).
14. Renu, G., Babylakshmi, M., Akhilesh, P. & Prasad, T. S. K. Human Protein Reference Database–2009 update. *Nucleic Acids Res.* **37,** D767–D772 (2009).
15. Andrea, F. *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* **41,** D808–D815 (2013).
16. Bader, G. D., Doron, B. & Hogue, C. W. V. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* **31,** 248–250 (2003).
17. Chen, H. *et al.* Pathway mapping and development of disease-pecific biomarkers: protein-based network biomarkers. *J. Cell. Mol. Med.* **19,** 297–314 (2015).
18. Yuan, L., Zheng, C. H., Xia, J. F. & Huang, D. S. Module Based Differential Coexpression Analysis Method for Type 2 Diabetes. *Biomed Res. Int*. **2015** (2015).
19. Didier, G., Brun, C. & Baudot, A. Identifying communities from multiplex biological networks. *Peerj* **3,** e1525 (2015).
20. Oh, K., Hwang, T., Cha, K. & Yi, G.-S. Disease association and inter-connectivity analysis of human brain specific co-expressed functional modules. *Biol. Res.* **48,** 67 (2015).
21. Sharma, A. *et al.* A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes. *Hum. Mol. Genet.* **24,** 3005–3020 (2015).
22. Liu, Y. *et al.* DiME: a scalable disease module identification algorithm with application to glioma progression. *Plos One* **9,** e86693 (2014).
23. Hormozdiari, F., Penn, O., Borenstein, E. & Eichler, E. The discovery of integrated gene networks for autism and related disorders. *Genome Res.* **25,** 142–154 (2014).
24. Zeng, T. *et al.* Deciphering early development of complex diseases by progressive module network. *Methods* **67,** 334–343 (2014).
25. Yang, R., Daigle, B. J., Petzold, L. R. & Doyle, F. J. Core module biomarker identification with network exploration for breast cancer metastasis. *BMC bioinformatics* **13,** 12 (2012).
26. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* **12,** 56–68 (2011).
27. Lin, C. C. *et al.* Dynamic functional modules in co-expressed protein interaction networks of dilated cardiomyopathy. *BMC Syst. Biol.* **4,** 138 (2010).
28. Baranzini, S. E. *et al.* Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Hum. Mol. Genet.* **18,** 2078–2090 (2009).
29. Wang, X., Dalkic, E., Wu, M. & Chan, C. Gene module level analysis: identification to networks and dynamics. *Curr. Opin. Biotechnol.* **19,** 482–491 (2008).
30. Guimera, R. & Amaral, L. A. N. Functional cartography of complex metabolic networks. *Nature* **433,** 895–900 (2005).
31. Ghiassian, S. D., Menche, J. & Barabási, A.-L. A DiseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human Interactome. *Plos Comput. Biol.* **11,** e1004120 (2015).
32. Menche, J. *et al.* Uncovering disease-disease relationships through the incomplete interactome. *Science* **347,** 1257601 (2015).
33. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452,** 423–428 (2008).
34. Han, J.-D. J., Dupuy, D., Bertin, N., Cusick, M. E. & Vidal, M. Effect of sampling on topology predictions of protein-protein interaction networks. *Nat. Biotechnol.* **23,** 839–844 (2005).
35. Dennis, G. Jr. *et al.* DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.* **4,** P3 (2003).
36. Lorenz, W. W. *et al.* Microarray analysis and scale-free gene networks identify candidate regulators in drought-stressed roots of loblolly pine (P. taeda L.). *BMC Genomics* **12,** 264 (2011).
37. Kazi, J. U. & Rönnstrand, L. Suppressor of cytokine signaling 2 (SOCS2) associates with FLT3 and negatively regulates downstream signaling. *Mol. Oncol.* **7,** 693–703 (2013).

38. Placke, T. *et al.* Requirement for CDK6 in MLL-rearranged acute myeloid leukemia. *Blood* **124,** 13–23 (2014).
39. Ward, I. M., Minn, K., van Deursen, J. & Chen, J. p53 Binding protein 53BP1 is required for DNA damage responses and tumor suppression in mice. *Mol. cell. Biol.* **23,** 2556–2563 (2003).
40. LePage, G. Phosphorylated Intermediates in Tumor Glycolysis I. Analysis of Tumors. *Cancer Res.* **8,** 193–196 (1948).
41. Solaini, G., Sgarbi, G. & Baracca, A. Oxidative phosphorylation in cancer cells. *Biochim. Biophys. Acta.* **1807,** 534–542 (2011).
42. Sven, G., Nora, A., Carsten, H., S Gail, E. & Serkova, N. J. Imatinib (STI571)-mediated changes in glucose metabolism in human leukemia BCR-ABL-positive cells. *Clin. Cancer Res.* **10,** 6661–6668 (2004).
43. Hu, N., Gibo, D. M. & Debinski, W. Cytokine up-regulation of IL13Ra2 in GBM cells leads to an increased potency of recombinant IL13 cytotoxin. *Cancer Ther*. **3,** 531–542 (2005).
44. Götze, S., Wolter, M., Reifenberger, G., Müller, O. & Sievers, S. Frequent promoter hypermethylation of Wnt pathway inhibitor genes in malignant astrocytic gliomas. *Int. J. Cancer* **126,** 2584–2593 (2010).
45. Lesnick, T. G. *et al.* A genomic pathway approach to a complex disease: axon guidance and Parkinson disease. *Plos Genet.* **3,** e98 (2007).
46. Berridge, M. J. Calcium signalling and psychiatric disease: bipolar disorder and schizophrenia. *Cell Tissue Res.* **357,** 477–492 (2014).
47. Asgharzadeh, S. *et al.* Prognostic significance of gene expression profiles of metastatic neuroblastomas lacking MYCN gene amplification. *J. Natl. Cancer Inst.* **98,** 1193–1203 (2006).
48. Sumimoto, H., Imabayashi, F., Iwata, T. & Kawakami, Y. The BRAF–MAPK signaling pathway is essential for cancer-immune evasion in human melanoma cells. *J. Exp. Med.* **203,** 1651–1656 (2006).
49. Behrens, J. The role of cell adhesion molecules in cancer invasion and metastasis. *Breast Cancer Res. Treat.* **24,** 175–184 (1993).
50. Su, S.-Y. *et al.* Transcriptomic analysis of EGb 761-regulated neuroactive receptor pathway *in vivo*. *J. Ethnopharmacol.* **123,** 68–73 (2009).
51. Turrin, N. P. & Plata-Salamán, C. R. Cytokine–cytokine interactions and the brain. *Brain Res. Bull.* **51,** 3–9 (2000).
52. Chowdhury, M. A. *et al.* Identification of novel functional sequence variants in the gene for peptidase inhibitor 3. *BMC Med. Genet.* **7,** 49 (2006).
53. Mohanan, V., Temburni, M. K. & Kappes, J. C. & Galileo, D. S. L1CAM stimulates glioma cell motility and proliferation through the fibroblast growth factor receptor. *Clin. Exp. Metastasis* **30,** 507–520 (2013).
54. Mocellin, S. & Nitti, D. TNF and cancer: the two sides of the coin. *Front. Biosci.* **13,** 2774–2783 (2008).
55. Li, J. *et al.* Identification of high-quality cancer prognostic markers and metastasis network modules. *Nat. Commun.* **1,** 34 (2010).
56. Wang, E. Understanding genomic alterations in cancer genomes using an integrative network approach. *Cancer Lett.* **340,** 261–269 (2013).
57. Cui, Q. *et al.* A map of human cancer signaling. *Mol. Syst. Biol.* **3,** 152 (2007).
58. Zaman, N. *et al.* Signaling network assessment of mutations and copy number variations predict breast cancer subtype-specific drug targets. *Cell Rep*. **5,** 216–223 (2013).
59. Wang, E. *et al.* Predictive genomics: A cancer hallmark network framework for predicting tumor clinical phenotypes using genome sequencing data. *Semin. Cancer Biol.* **30,** 4–12 (2015).
60. Phillips, H. S. *et al.* Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell* **9,** 157–173 (2006).
61. Cai, Y., Fendler, B. & Atwal, G. S. Utilizing RNA-seq data for cancer network inference. *Genomic Signal Processing and Statistics,(GENSIPS), 2012 IEEE International Workshop on*. 46–49 (2012).
62. Friedman, J., Hastie, T. & Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9,** 432–441 (2008).
63. Menéndez, P., Kourmpetis, Y. A., ter Braak, C. J. & van Eeuwijk, F. A. Gene Regulatory Networks from Multifactorial Perturbations Using Graphical Lasso: Application to the DREAM4 Challenge. *Plos One* **5,** e14147 (2010).
64. Botstein, D. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25,** 25–29 (2000).
65. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4,** 44–57 (2008).
66. Doolittle, J. M. & Gomez, S. M. Mapping protein interactions between Dengue virus and its human and insect hosts. *Plos Negl. Trop. Dis*. **5,** e954 (2011).
67. Chen, J. *et al.* Pathway enrichment analysis of human osteosarcoma U-2 OS bone cells expose to dexamethasone. *Eur. Rev. Med. Pharmacol. Sci.* **18,** 2631–2639 (2014).
68. Jing, R., Sun, J., Wang, Y., Li, M. & Pu, X. PML: A parallel machine learning toolbox for data classification and regression. *Chemometr. Intell. Lab*. **138,** 1–6 (2014).
69. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. *Mach. Learn.* **46,** 389–422 (2002).
70. Vapnik, V. N. An overview of statistical learning theory. *IEEE Trans. on Neural Netw*. **10,** 988–999 (1999).
71. Chang, ChihChung, Lin & ChihJen. LIBSVM: A library for support vector machines. *ACM Transactions on TIST* **2,** 389–396 (2011).

Acknowledgements

This work was supported by the grants from the National Natural Science Foundation of China (No. 21575094, No. 21375090) and the National High Technology Research and Development Program of China (No. 2015AA020104).

Author Contributions

Z.W. and M.L. conceived and designed the experiments, J.X. performed the experiments, J.X., R.J. and Y.D. contributed analysis tool, J.X., Z.W. and M.L. wrote the paper. R.J. and Y.L. designed the web service used in analysis. All authors discussed the results and commented on the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Xu, J. *et al.* A new strategy for exploring the hierarchical structure of cancers by adaptively partitioning functional modules from gene expression network. *Sci. Rep.* **6**, 28720; doi: 10.1038/srep28720 (2016).