Cancer-Risk Module Identification and Module-Based Disease Risk Evaluation: A Case Study on Lung Cancer

Xu Jia[9], Zhengqiang Miao[9], Wan Li, Liangcai Zhang, Chenchen Feng, Yuehan He, Xiaoman Bi, Liqiang Wang, Youwen Du, Min Hou, Dapeng Hao, Yun Xiao, Lina Chen*, Kongning Li*

College of Bioinformatics Science and Technology, Harbin Medical University,Harbin,Hei Longjiang Province, China

Abstract

Gene expression profiles have drawn broad attention in deciphering the pathogenesis of human cancers. Cancer-related gene modules could be identified in co-expression networks and be applied to facilitate cancer research and clinical diagnosis. In this paper, a new method was proposed to identify lung cancer-risk modules and evaluate the module-based disease risks of samples. The results showed that thirty one cancer-risk modules were closely related to the lung cancer genes at the functional level and interactional level, indicating that these modules and genes might synergistically lead to the occurrence of lung cancer. Our method was proved to have good robustness by evaluating the disease risk of samples in eight cancer expression profiles (four for lung cancer and four for other cancers), and had better performance than the WGCNA method. This method could provide assistance to the diagnosis and treatment of cancers and a new clue for explaining cancer mechanisms.

Citation: Jia X, Miao Z, Li W, Zhang L, Feng C, et al. (2014) Cancer-Risk Module Identification and Module-Based Disease Risk Evaluation: A Case Study on Lung Cancer. PLoS ONE 9(3): e92395. doi:10.1371/journal.pone.0092395

Editor: Ying Xu, University of Georgia, United States of America

Received July 12, 2013; **Accepted** February 21, 2014; **Published** March 18, 2014

Funding: Funding provided by the National Natural Science Foundation of China (No. 61272388 and No. 31301040); Oversea Scholars Project funded by Education Department of Heilongjiang Province (NO. 1155H012); and the Master Innovation Funds of Heilongjiang Province (No. YJSCX2012-209HLJ and YJSCX2012-224HLJ). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: chenlina@ems.hrbmu.edu.cn (LC); kongningli@hotmail.com (KL)

[9] These authors contributed equally to this work.

Introduction

Cancer is caused by aberration of multiple genes, and thus its pathogenesis is very complex and inconclusive [1,2,3]. Cancer-related genes possess diverse functions [4,5], while genes with similar functions are likely to be co-expressed [6,7] and located in neighboring areas (known as network modules) [8,9] in biological networks. The modules reveal the mechanism of multiple genes underlying the disease and evaluate the risk of the disease. Effective identification of cancer risk modules can assist cancer researches [10,11,12,13].

Disease risk of cancer-related modules calculated from a specific biological background can be a significant measure for clinical prediction of cancer diagnosis [14,15,16,17,18]. Several computational approaches have been developed for the disease risk module analysis, including detection of differentially correlated gene clusters and gene-specific analysis based the co-expression network [19,20,21,22]. For example, weighted gene co-expression network analysis (WGCNA) is a mature technique and identifies gene modules as candidate biomarkers or therapeutic targets based on the co-expression network [23,24]. WGCNA has been used to study complex diseases, such as metabolic syndrome [25], schizophrenia [26], and heart failure [27]. The expression activities of disease risk modules were (induced or repressed) different among clinical conditions (in tumor progress)[14].

Furthermore, it is feasible to identify cancer risk modules from co-expression networks using network-based methods. The analysis of gene co-expression networks shows that genes within the same modules appear to have similar expression patterns, share common regulatory mechanisms [28,29,30], and thus have strong associations with specific biological functions that determine the behaviors or phenotypes of cells [31,32]. Modules derived from co-expression network were organized into a higher-order structure correlated with clinical characteristics, which provided insights into the underlying biology of glioma [33]. Four modules of ovarian cancer from a co-expression network were distinguished to be significantly associated with biological processes such as cell cycle and DNA replication in Gene Ontology (GO) categories[34]. The co-expression modules associated with T-helper differentiation and TGF-beta pathways improved clinical outcome of hormone-insensitive breast cancers after treatment [35]. Moreover, sample signatures/labels considered in evaluation of cancer-related risk modules would offer a new clue for revealing the mechanisms of diseases [36]. Researches have revealed that it is necessary to explore the relationships between gene functions and disease risks [37,38]. The co-expression networks taking into account of biological functions would be more robust and authentic [39,40], and the modules obtained from these networks could better reflect the function information of the diseases.

In this paper, a new method was proposed to identify cancer-risk modules and evaluate the module-based disease risks of samples. A highly-confident co-expression network with functional similarity information was first constructed by using

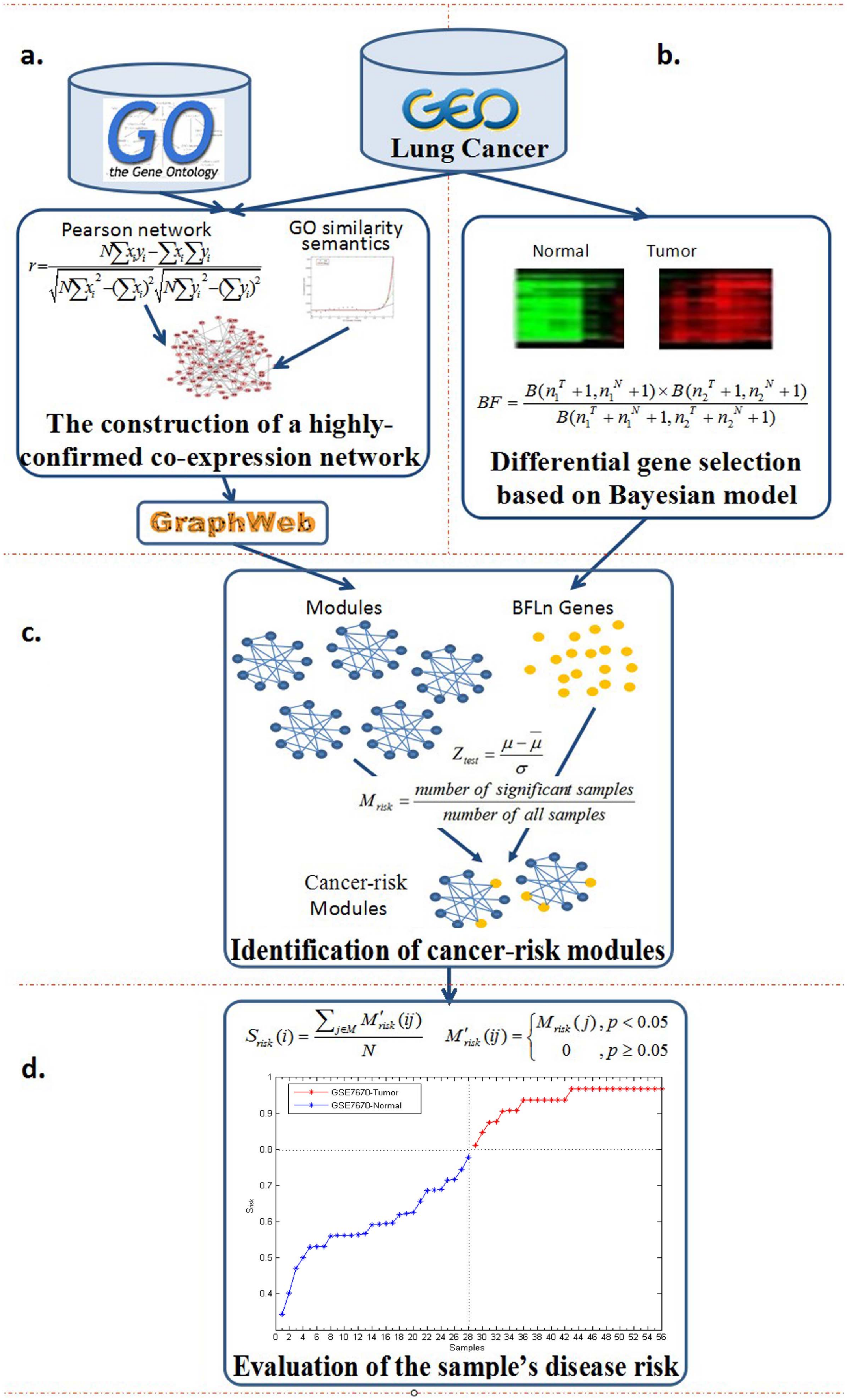

a.
GO
the Gene Ontology
GEO
Lung Cancer
b.
Pearson network
$r=\frac{N\sum x_i y_i-\sum x_i\sum y_i}{\sqrt{N\sum x_i^2-(\sum x_i)^2}\sqrt{N\sum y_i^2-(\sum y_i)^2}}$
GO similarity semantics
The construction of a highly-confirmed co-expression network
GraphWeb
Normal
Tumor
$BF=\frac{B(n_1^T+1,n_1^N+1)\times B(n_2^T+1,n_2^N+1)}{B(n_1^T+n_1^N+1,n_2^T+n_2^N+1)}$
Differential gene selection based on Bayesian model
c.
Modules
BFLn Genes
$Z_{test}=\frac{\mu-\bar{\mu}}{\sigma}$
$M_{risk}=\frac{number\ of\ significant\ samples}{number\ of\ all\ samples}$
Cancer-risk Modules
Identification of cancer-risk modules
d.
$S_{risk}(i)=\frac{\sum_{j\in M}M'_{risk}(ij)}{N}$
$M'_{risk}(ij)=\begin{cases}M_{risk}(j), & p<0.05\\ 0, & p\geq 0.05\end{cases}$
GSE7670-Tumor
GSE7670-Normal
1
0.9
0.8
0.7
0.6
0.5
0.4
Srisk
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56
Samples
Evaluation of the sample's disease risk

Figure 1. Cancer-risk Modules Identification and Module-based Disease risk Evaluation.
doi:10.1371/journal.pone.0092395.g001

Table 1. The number of the tumor samples and the normal samples in the expression profiles.

	GSE10072	GSE21933	GSE27262	GSE40791	GSE14520	GSE15781	GSE20437	GSE26126
Cancer	Lung Cancer				Liver Cancer	Colon Cancer	Breast Cancer	Prostate Cancer
GPL	GPL96	GPL6254	GPL570	GPL570	GPL3921	GPL2986	GPL96	GPL8490
Tumor	58	21	25	94	64	13	18	181
Normal	49	21	25	100	64	10	15	12

doi:10.1371/journal.pone.0092395.t001

expression profiles in lung cancer, and then candidate modules were identified. The cancer risks of the modules were scored by introducing sample labels, then the significant cancer-risk modules were screened out by randomized trials. Finally, the disease risks of samples were evaluated based on the cancer-risk modules. These modules were expected to provide evidence for disease diagnosis, treatment and clinical analysis in the future. Identification of cancer-risk modules and evaluation of module-based disease risks were performed in the following steps (Figure 1).

Materials and Methods

Materials

Cancer gene expression data were obtained from the Gene Expression Omnibus(GEO, http://www.ncbi.nlm.nih.gov/geo/)[37]. Here, our research was based on the profile GSE7670 [41]in GPL96 including 20,995 genes of 56 samples (28 lung cancer patients and 28 normal controls), for which patients underwent surgery for lung cancer at the Taipei Veterans General Hospital. These expression profiles (GSE10072, GSE21933, GSE27262, GSE40791, GSE14520, GSE15781, GSE20437, GSE26126) (Table 1) with disease and normal samples were used to analyze the robustness of our method and compare with the WGCNA method. Gene function information was obtained from Gene Ontology (GO, http://www.geneontology.org/) [42], updated to May 2011. Protein interaction information (95537 high-confidence interactions between 12359 genes) was downloaded from iRefWeb (http://www.wodaklab.org/iRefWeb/) [43], updated to April 13, 2012 of the 9th version. The information of 1824 protein complexes was obtained from Munich Information Center for Protein Sequences (MIPS, http://mips.helmholtz-muenchen.de/genre/proj/corum, Corum Release February 2012 available).

a. The construction of a highly confident co-expression network. A method was introduced to create a highly confident co-expression network by taking both co-expression correlation and functional similarity. This method was performed as follows:

First, the Pearson correlation coefficient [44] r was used to represent the co-expression relationship between every pair of genes and calculated as follows:

$$r=\frac{N\sum x_i y_i-\sum x_i\sum y_i}{\sqrt{N\sum x_i{}^2-\left(\sum x_i\right)^2}-\sqrt{N\sum y_i{}^2-\left(\sum y_i\right)^2}}$$

where N is the number of samples in an expression profile, x_i and y_i are the expression levels of genes x and y in the i-th sample.

Second, GO semantic similarity was used to represent the functional similarity between every pair of genes [45].

(1) The similarity score of GO term A was defined as:

$$S_{GO}(A)=\sum_{t\in T_A} S_A(t)$$

Table 2. The number of samples (tumor/normal and high/low expression) for one gene.

	+	−	Total
T	n_1^T	n_2^T	$n_1^T + n_2^T$
N	n_1^N	n_2^N	$n_1^N + n_2^N$
Total	$n_1^T + n_1^N$	$n_2^T + n_2^N$	$n_1^T + n_2^T + n_2^N + n_2^N$

T represents tumor samples and N for normal ones, and "+" stands for high expression (above-average) and "-"for low expression (below-average). n_1^T and n_2^T refers to the number of tumor samples with high expression and low expression, and n_1^N and n_2^N for the number of normal ones with high expression and low expression.

doi:10.1371/journal.pone.0092395.t002

$$S_A(t)=\begin{cases} 1 & , t=A \\ \max\{W_e \times S_A(t')|t'\in childrenof(t)\} & , t\neq A \end{cases}$$

$$W_e=\begin{cases} 0.8 & e\ is\ "is\ a"\ edge \\ 0.6 & e\ is\ "part\ of"\ edge \end{cases}$$

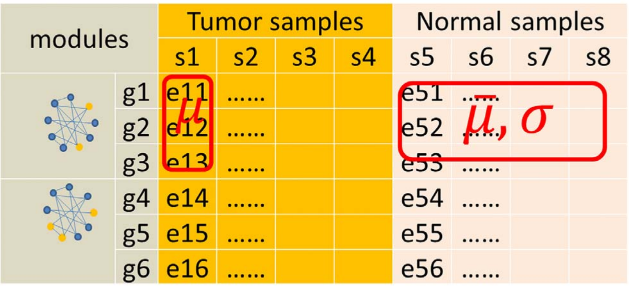
where T_A includes term A and all its parent terms; W_e is the weight of edge; and it is 0.8 for 'is-a' relationship and 0.6 for 'part-of' relationship.

(2) The semantic similarity between term A and term B, $S_{GO2GO}(A,B)$, was calculated as follows:

$$S_{GO2GO}(A,B)=\frac{\sum\limits_{t\in T_1\cap T_2}(S_A(t)+S_B(t))}{S_{GO}(A)+S_{GO}(B)}$$

A gene's functions were considered as a set of GO terms in Gene Ontology. Thus, functions of genes G1 and G2 corresponded to GO sets $GO_1=\{go_{11},go_{12},\ldots,go_{1m}\}andGO_1=\{go_{21},go_{22},\ldots,go_{2n}\}$, m and n are the number of terms in GO1 and GO2 respectively.

(3) The semantic similarity between G1 and G2 was defined as:



modules		Tumor samples				Normal samples			
		s1	s2	s3	s4	s5	s6	s7	s8
	g1	e11	......			e51	......		
	g2	e12	......			e52	......		
	g3	e13	......			e53	......		
	g4	e14	......			e54	......		
	g5	e15	......			e55	......		
	g6	e16	......			e56	......		

Figure 2. Z-test. Where μ means the average expression value of all genes in module1 for the tumor sample s1; e11 is the expression value of g1 in module1 for s1, so do others; $\bar{\mu}$ means the average expression value of all genes for all normal samples; σ is the standard deviation of all normal samples.

doi:10.1371/journal.pone.0092395.g002

$$S_{G2G}(G1,G2)=\frac{\sum\limits_{1\leq i\leq m}\max_{1\leq j\leq n}(S_{GO2GO}(go_{1i},go_{2j}))+\sum\limits_{1\leq i\leq n}\max_{1\leq j\leq m}(S_{GO2GO}(go_{2i},go_{1j}))}{m+n}$$

The robust gene pairs were retained by the function similarity. Therefore, a highly-confident co-expression network was constructed by analyzing the Pearson correlation coefficient and GO semantic similarity.

b. Differential gene selection based on Bayesian model. A Bayesian model [46,47] was used to screen the differential genes. Bayesian approaches compare the probability of an association between a gene expression and a disease to the probability given no such association. The formula was as follows:

$$BFLn=$$
$$BLn(n_1^T+1,n_1^N+1)+BLn(n_2^T+1,n_2^N+1)-BLn(n_1^T+n_1^N+1,n_2^T+n_2^N+1)$$

where $n_1{}^T$, $n_2{}^T$, $n_2{}^N$ and $n_2{}^N$ are the number of samples (tumor/normal and high/low expression) for one gene (Table 2). B denotes the Beta function, defined by

$$B(n^T+1,n^N+1)=\frac{n^T!n^N!}{(n^T+n^N+1)!}$$

BLn is the log value of B.

When $BFLn>0$, there was relationship between a disease and gene expression; when $BFLn<0$, no relationship.

A randomized test was designed to calculate the significance of $BFLn$ by stochastically disturbing $n_1{}^T$, $n_2{}^T$, $n_2{}^N$ and $n_2{}^N$ and retaining stable sum; after 10,000 times, the p-value was the proportion when the random $BFLn$ was larger than the real value. Genes with $p<0.05$ were selected as differentially expressed genes (DE-genes).

c. Identification of cancer-risk modules. The online module mining tool GraphWeb (http://biit.cs.ut.ee/graphweb/) [48] was chosen to find co-expression modules. GraphWeb is designed to analyze individual or multiple merged networks, search for conserved features across multiple species, mine large biological networks for smaller modules, and compare results of high-throughput datasets. Markov Cluster (MCL) [49] algorithm via the GraphWeb tool was applied to prune the network and to find gene modules. The MCL algorithm simulates a stochastic flow in the expression

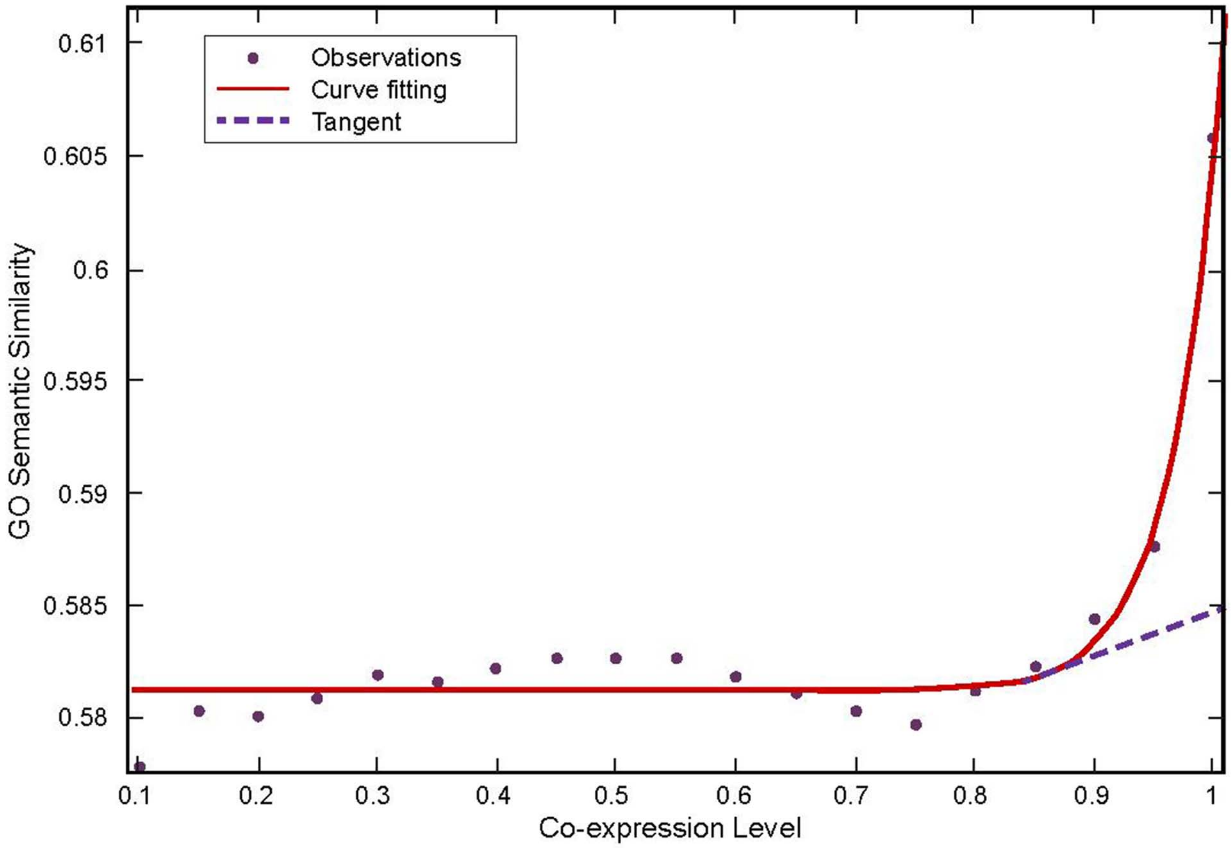


Figure 3. Co-expression Level and GO Semantic Similarity. Purple point means observations, red line indicates the curve fitting, the dotted curve represents the first order tangent.
doi:10.1371/journal.pone.0092395.g003

graph and removes edges that are visited infrequently, resulting in a collection of densely connected groups of genes. The parameter of Markov clustering parameter was set to a default value 1.8.

The candidate modules containing the DE-genes were selected to evaluate the disease risks. Next, *Z*-test[50] was applied to assess the relationship between individual tumor samples and modules (Figure 2).

$$Z_{test} = \frac{\mu - \overline{\mu}}{\sigma}$$

Finally, the significant samples with Z-test higher than the significance threshold ($\alpha = 0.05$) were picked out. To measure the risk of each module, we defined:

$$M_{risk} = \frac{number\ of\ significant\ samples}{number\ of\ all\ samples}$$

M_{risk} could be used to assess the disease risk of a candidate module. For each candidate module, 10,000 random modules were constructed by randomly selecting genes from the background gene set with equal numbers of module genes. Then, M_{risk} was calculated for each random module, and the proportion of modules with M_{risk} larger than the real value (the significance *p*-value) was computed. Modules with $p<0.05$ were considered as cancer-risk modules.

d. Evaluation of the sample's disease risk. To evaluate the module-based disease risk of each sample, we defined:

$$S_{risk}(i) = \frac{\sum_{j \in M} M'_{risk}(ij)}{N}$$

$$M'_{risk}(ij) = \begin{cases} M_{risk}(j)\ , p < 0.05 \\ 0\ \ \ \ \ \ , p \geq 0.05 \end{cases}$$

where M includes all cancer-risk modules, N is the number of cancer-risk modules, $M'_{risk}(ij)$ means the cancer-risk of the sample i about the module j, and p is the significance of Z-test.

Cancer-risk modules were applied to evaluate samples by calculating the module-based disease risk of each sample. Then evaluation performance was estimated by a receiver operating characteristic (ROC) curve.

Table 3. Lung cancer-risk modules.

Risk	ID	Size	Genes	M_{risk}	p-value
high	M2	171	ZEB1, CAV1, HYAL2, MMP12, CLU, TIMP3, DKK3, LPL, TCF21, FOXF1…	1	0.0043
	M72	9	ASPM*,BUB1B,CCNB2,CEP55,KPNA2*,MAD2L1,PBK,TPX2,TRIP13	1	0.0036
	M46	13	BARD1,CDT1,DLGAP5*,DONSON*,GINS1,KIF4A*, MCM7*,MCM3,MLF1IP*,NDC80,PAQR4,TMEM48,TTK	1	0.0062
	M39	14	ADRM1,BYSL,CKS1B,CRABP2,DNAJA3,HAX1,LSM12, MPZL1*,MRPL17,MRPS7,NME4,RPN2,SLC2A4RG,STRA13	1	0.0067
	M281	3	CRYAB*,HSPB2*,VGLL3*	1	0.0018
	M82	9	ALG3*,EIF2S1,HSPB11,LRRC42,MCTS1,P4HA2,PSMA5,SEC61G,VARS	1	0.004
	M61	11	ADAMTS8*,CSRP1*,KCNK3*,LINC00312*,MYH11*, MYLK,PDE2A,PKNOX2*,RASL12*,SETBP1,TACC1*	1	0.0058
	M266	3	CDCA3,GALNT6,IDH2*	1	0.0017
	M340	3	MRPS34,NUBP2,SNRNP25*	1	0.0015
	M363	3	DDR1*,FLAD1*,SPINT1	1	0.002
middle	M62	11	CCNB1,CKAP2,KIF11*,KIF20A*,MCM4,MELK*, NCAPG,NETO2*,PRC1*,SHCBP1,TOP2A*	0.9642	0.0187
	M27	17	CCT6A*,EIF2AK1*,EIF3B,FKBP14*,GART,GINS4,GNL3, HEATR2*,KLHL7*,LSM5*,MRPS17*,MRPS33*,PHLDA2, POLD2,PPP1R14B*,PSMD2,TMEM106B*	0.9642	0.0304
	M268	3	HPRT1*,SCRN1*,TPBG*	0.9642	0.0065
	M102	8	AVL9,CDK5,CORO1B,CHPF2*,ITPKA,NDUFS8,PPP1CA,SSH3	0.9642	0.0172
	M63	10	A2M*,CASP1*,CD97*,FABP4*,GAS6*,GMFG*, PDLIM2*,PLEKHO2*,RARRES2,TRPV2*	0.9642	0.0171
	M54	12	CLDN5*,CRIM1*,DOCK6,FGR*,ICAM2*,INPP1*, KANK3*,LIMS2*,LRRC32,PCDH12*,PTGIR*,RASIP1*	0.9642	0.0223
	M258	4	FZR1*,CLDN4,LY6E,PRSS8	0.9642	0.0076
	M188	5	BLVRA,KIAA0391,PSMA6,SRP54*,TFPI2	0.9642	0.0091
	M297	3	AHCY*,PKP3,SLC38A1	0.9642	0.0088
	M321	3	GLO1,EGFL7,PDXDC1*	0.9642	0.0056
	M180	5	DHTKD1*,MEA1,SLC35A2,TMED3*,TPMT	0.9642	0.0096
	M86	9	CDKL2,ENY2,HAND1,LY6D,ORM1,ORM2,RAB25*,S100G,TSTA3*	0.9642	0.0159
	M387	3	GALNTL2*,SAR1B*,TSPAN6*	0.9642	0.0062
low	M157	5	DHFR,DTL,GMPS,MYBL2,RFC4*	0.9285	0.0234
	M241	4	COG8,FAM158A,PDF,PSMB5*	0.9285	0.0207
	M249	4	KRT10,NIPSNAP1*,POLDIP2*,SEPHS2	0.9285	0.0173
	M314	3	FAM65A*,GIMAP5*,SEPP1*	0.9285	0.0159
	M280	3	GYPC,PTGDS*,RPL15	0.9285	0.016
	M144	6	BCKDK,DECR2,GALE,NDUFB11,PYCR1*,RRNAD1	0.9285	0.028
	M316	3	CTSA,ERGIC3,PAFAH1B3*	0.8928	0.0313

Risk is modules category, ID indicate the identifier of cancer-risk modules, size is the module scale, namely the number of genes in the module, genes is the genes in the modules and the genes which were marked * were DE-genes, M_{risk} is the cancer risk of modules, p-value is significance p value of random randomized test.

doi:10.1371/journal.pone.0092395.t003

Results

The highly-confident co-expression network

The Pearson correlation coefficient and the GO semantic similarity of every pair of genes in the expression profile GSE7670 were calculated. After that, curve fitting was applied to analyze the variation trend of average distribution of co-expression value with GO semantic similarity at a 0.05 interval (Figure 3). Functional similarity increased when co-expression level was over the tangency point. Therefore, the pairs of genes with functional similarity over 0.582 and Pearson correlation coefficient over 0.82 (the tangent point) were selected to create the highly-confident co-expression network, which consisted of 9841 nodes and 112,605 edges.

Cancer-risk Modules

A total of 472 DE-genes were screened out by applying BFLn to the expression profile GSE7670. Then 75 candidate disease

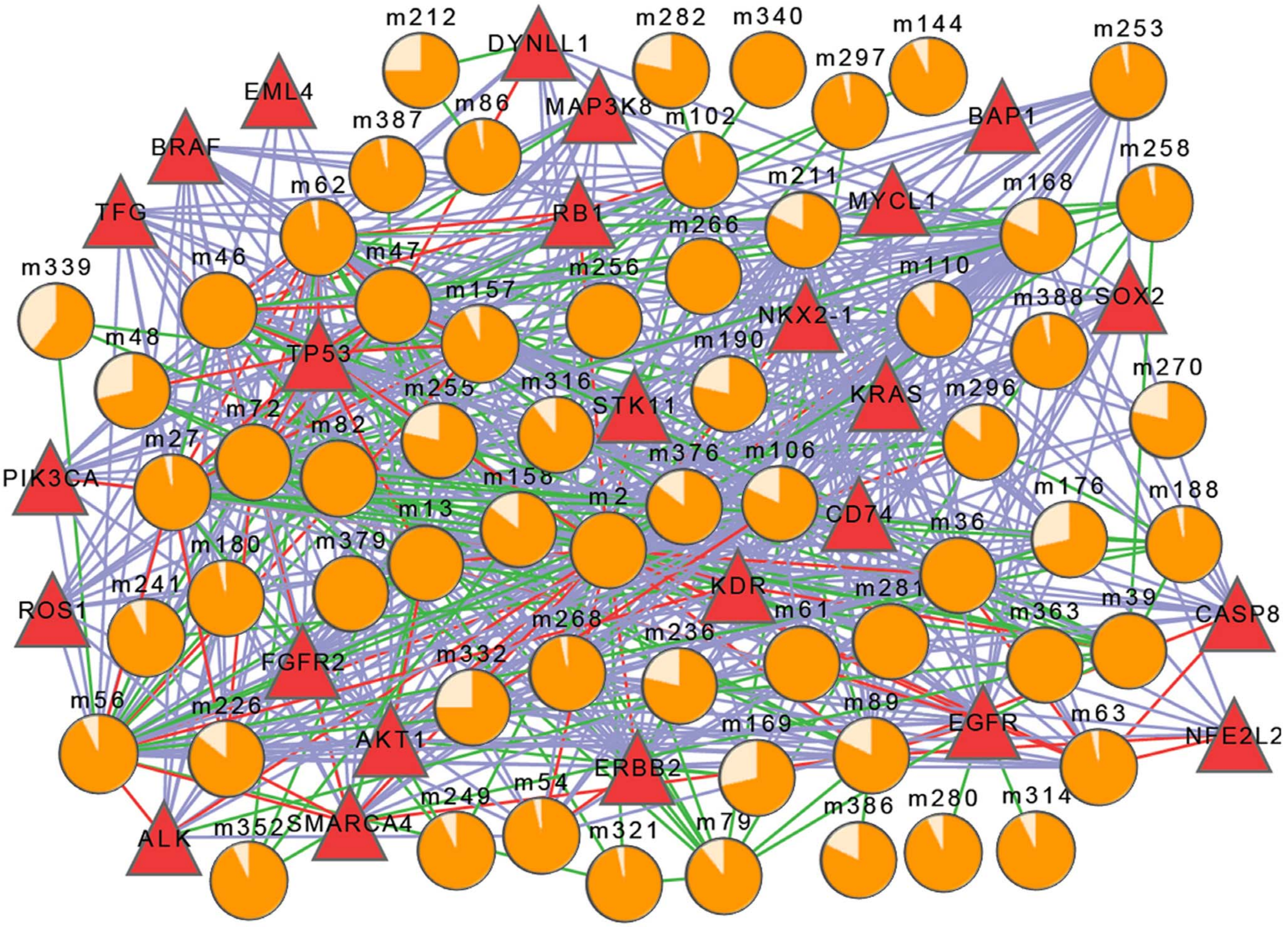


Figure 4. The relationship network of cancer-risk modules and lung cancer genes. The circles indicate cancer-risk modules, and the proportion of orange parts indicates cancer risk (M_{risk}). The disease-causing genes is represented by red triangles. Edges' colors indicate the relationships, purple represents for the protein-protein interaction, green for function sharing, and red for both functional and interaction relationship.
doi:10.1371/journal.pone.0092395.g004

modules containing DE-genes were obtained through GraphWeb. After the randomized test, 31 lung cancer-risk modules were obtained (Table 3).

Evaluation of cancer-risk modules

The cancer-risk modules were evaluated at the functional level and interactional level. On one hand, functional enrichment was performed for each lung cancer-risk module using an online tool DAVID (http://david.abcc.ncifcrf.gov/home.jsp) [51], and then significantly enriched GO terms of each module were obtained (More modules are in the Table S1). On the other hand, the interactional relationships of modules were assessed using protein interaction data from iRefWeb. The relationship network of cancer-risk modules and known lung cancer genes was constructed on the basis of functional and interactional relationships (Figure 4). The results showed that lung cancer-risk modules were closely related with the lung cancer genes, which indicated that these modules and genes might synergistically cause lung cancer. For instance, m46 was associated with cell cycle regulation and phosphorylation [52], cell proliferation and cell cycle checkpoint

Table 4. Average Degree for three types of cancer-risk modules.

Risk	D_W	D_M	D_D	D_P	D_F	D_B
High	13.00	5.222	7.78	6.33	9.22	2.50
Middle	10.67	3.75	6.92	3.50	8.58	1.42
Low	4.71	2.00	2.70	2.14	2.80	0.28

D_W stands for degree of whole net, D_M for degree only between modules, D_D for degree only considered of modules with disease-causing genes, D_P for degree of the protein interaction edges(purple edges), D_F for degree of function edges(green edges), D_B for degree of both protein interaction and function(red edges).
doi:10.1371/journal.pone.0092395.t004

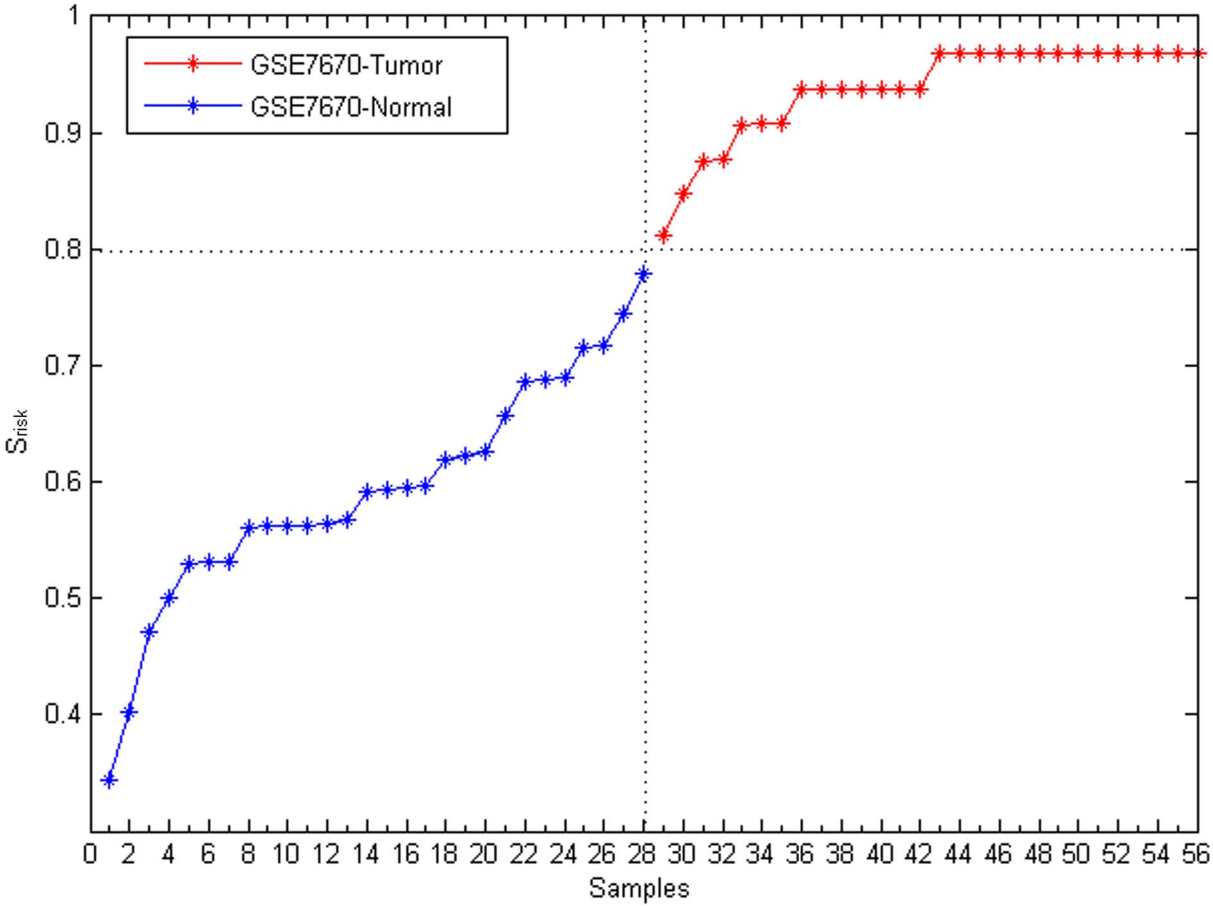


Figure 5. The lung cancer risk of each sample in GSE7670. X-axis is samples. Y-axis is the lung cancer risk score of individual samples, and it is ranked from smallest to largest. Red represents lung cancer samples; and blue represents normal samples.
doi:10.1371/journal.pone.0092395.g005

[53], and ATP binding [54] by interacting with known lung cancer genes KRAS, KDR and TP53, respectively. These functions were confirmed to be related to the occurrence of lung cancer. Another module m63 was significantly enriched in functions associated with the cancer, e.g. the response to corticosteroid stimulus, the response to organic substance, and glucocorticoid stimulus and steroid hormone stimulus together by interacting with known lung cancer genes KRAS, NFE2L2 and NKX2, respectively [55,56,57].

To further analyze the relationship network, the cancer-risk modules were classified into three types according to the risks: the high, middle and low risk modules (Table 3), and the corresponding degree distributions were calculated (Table 4). The results showed that the high risk modules tend to have high degrees. Namely, they had more connections with other modules and known disease genes at the functional and interactional levels. They played pivotal roles in the network.

Evaluation of the module-based disease risk

The lung cancer risk of each sample was evaluated by considering the cancer-risk modules. By measuring the lung cancer risk (S_{risk}), every sample in GSE7670 was evaluated. It turned out that every sample could be successfully identified as disease (S_{risk}>0.8) or normal (S_{risk}<0.8) based on its disease risk (Figure 5).

The robustness of our method

In order to verify the robustness of this method, first, other four expression profiles (GSE10072 from GPL96, the same as GSE7670; GSE27262 and GSE40791 from GPL570; and GSE21933 from GPL6254) about lung cancer and normal were evaluated, respectively (Table 1). The results showed that the module-based disease risks of cancer samples were higher than those of normal ones (Figure 6a). ROC curves were then plotted and the AUC values (>0.97) were used to measure the evaluation performances of the cancer-risk modules which were obtained by our method (Figure 6c). The method had good performance in the expression profiles not only from the same platform, but also from different platforms.

Next, we identified risk modules of liver cancer (GSE14520), colon cancer (GSE15781), breast cancer (GSE20437), and prostate cancer (GSE26126) in the same way, respectively (More cancer-risk modules information in the four cancers are in the Table S3). The cancer-risk modules were used to evaluate the disease risks of the samples, and the corresponding ROC curves were drawn (Figure 7).

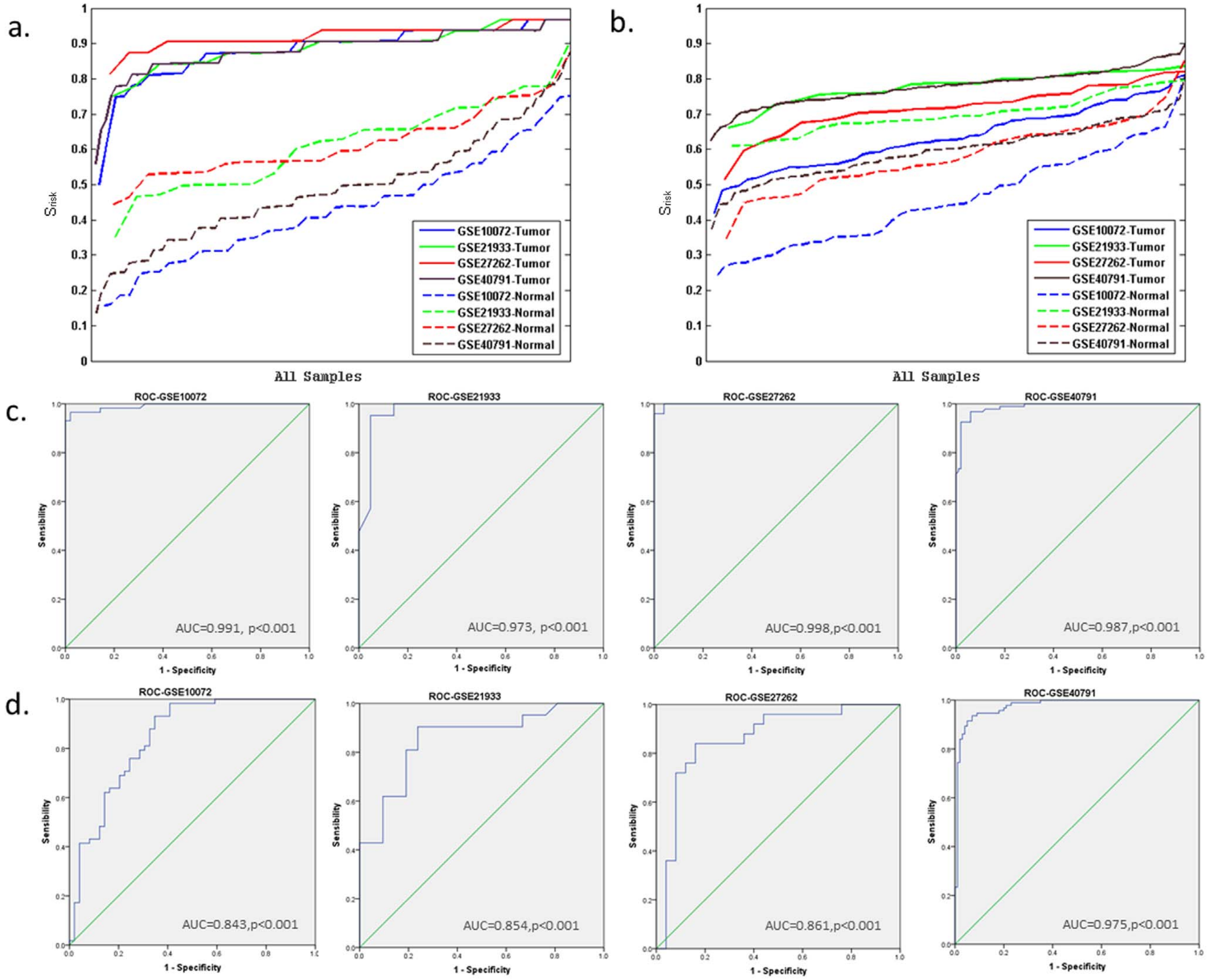


Figure 6. The robustness of our method and comparison with the WGCNA method. a) X-axis is samples. Y-axis is the lung cancer risk score of individual samples using our method, and it is ranked from the smallest to the largest. Blue represents GSE10072; green represents GSE21933; red represents GSE27262; and brown represents GSE4079. Full lines represent lung cancer samples; and dashed lines represent normal samples. The different experiment data sets have different numbers of the normal samples and the disease samples. In order to show the disease risk of every sample in four expression profiles intuitively, all samples of each expression profiles are distributed uniformly throughout x-axis. **b**) The figure is plotted the same way as a). The lung cancer risk of each sample is evaluated by the WGCNA method. **c**) Receiver operator characteristic curve using our method for the four lung cancer expression profiles (see Figure 7a). The areas under curve provided at lower right of each diagram. **d**) Receiver operator characteristic curve using the WGCNA method for the four lung cancer expression profiles (see Figure 7b).
doi:10.1371/journal.pone.0092395.g006

Method comparisons

The WGCNA method [24] is a widely used technique to construct gene modules within a network based on gene co-expression relationships. In this paper, the accuracy and robustness of WGCNA and our method were compared. Fifty seven lung cancer risk modules were obtained from GSE7670 using the WGCNA method. The lung cancer risk of every sample in GSE7670 itself was evaluated with the modules. Cancer risks of some cancer samples were smaller than those of normal ones (Figure 8), which indicated the WGCNA method could not completely identified samples as disease or normal as accurately, while our method could (Figure 5).

Then the evaluation of the samples' lung cancer risks was extended to other four expression profiles about lung cancer and normal (Figure 6b). It was found that the cancer risks of cancer samples were not significantly different from those of normal ones. The ROC curves were then used to evaluate the performance of the WGCNA method (Figure 6d). We found that our method had better accuracy and robustness than the WGCNA method (Figure 6).

Discussion

Studying the mechanisms of diseases by analyzing gene expression profiles appears to be a convenient and effective way. Considering the functional similarity could better reflect the function information of the disease. In this paper, a new method was proposed to identify thirty one cancer-risk modules and evaluate the module-based disease risks of samples by using a co-

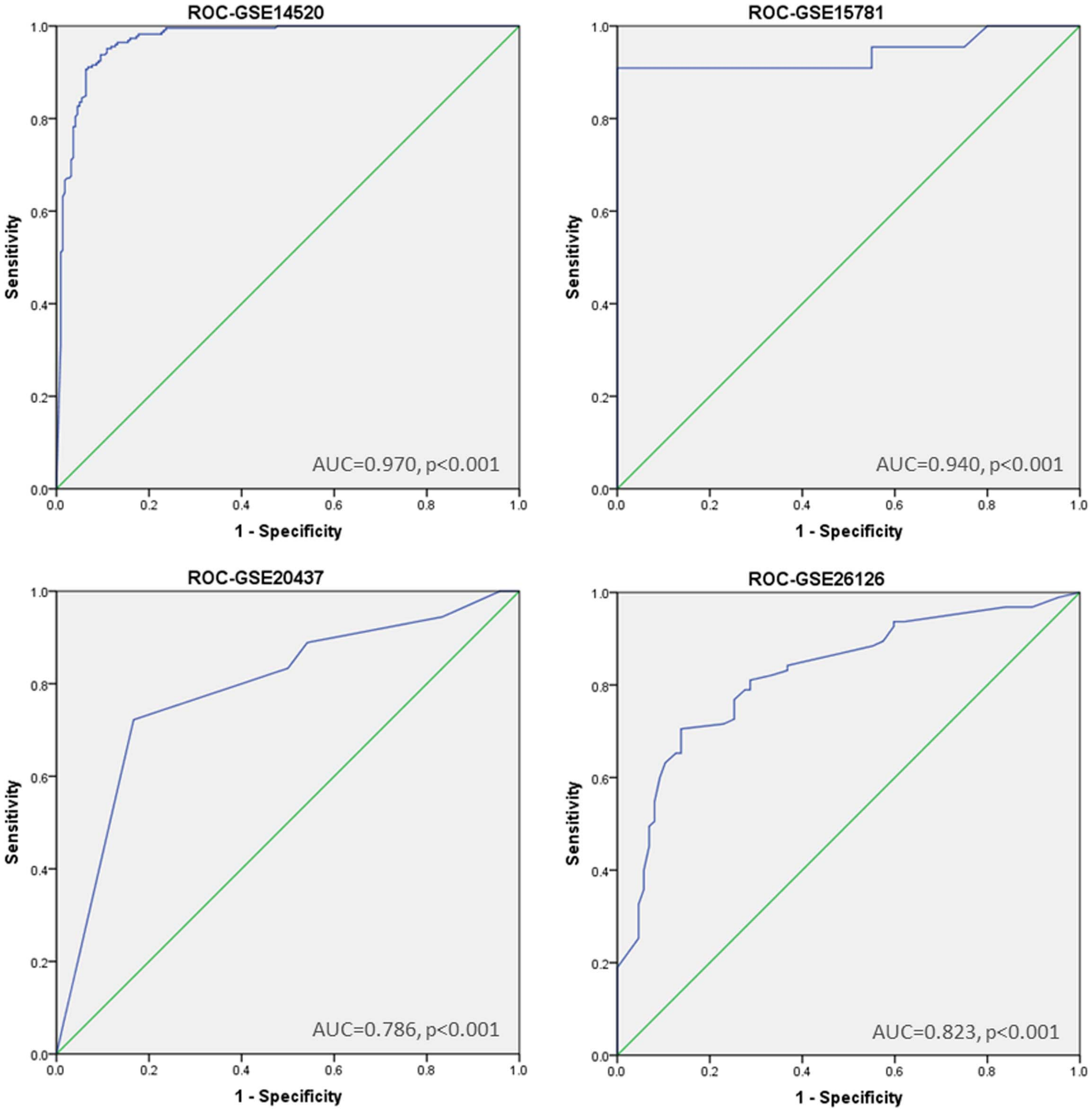


Figure 7. Receiver operator characteristic curve for expression profiles of liver cancer (GSE14520), colon cancer (GSE15781), breast cancer (GSE20437), and prostate cancer (GSE26126).
doi:10.1371/journal.pone.0092395.g007

expression network with functional similarity information. Finally, the relationship network of cancer-risk modules and cancer genes was constructed on the functional level and interactional level.

These modules were found to be closely related to cancers in the aspects of functions, interactions, and literature. Our method was proved to be fairly robust by evaluating the disease risks of samples in four lung cancer expression profiles and in four other cancers, and had better performance than the WGCNA method.

Cancer-risk modules and the evaluation of the module-based disease risk from this study were confirmed to be credible with the following considerations. (i) Differentially expressed genes were selected by using the BFLn method, which considered both gene expression and sample label distribution so as to eliminate outliers caused by bias expression of individual gene or experiment errors. (ii) Our gene network was of high confidence, because the method was used to calculate not only the co-expression correlation, but also functional similarities between genes. The gene pairs with both high expression

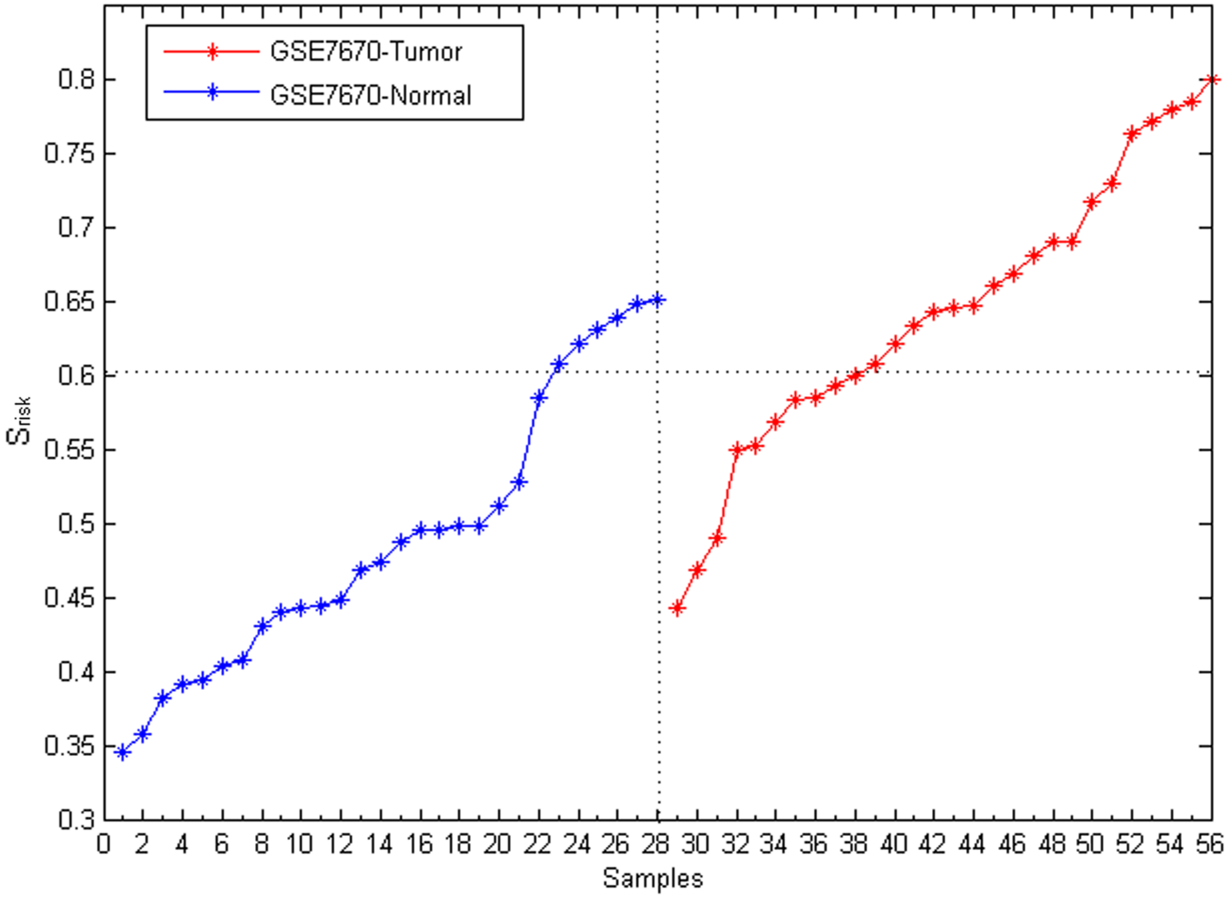


Figure 8. The lung cancer risk of each sample in GSE7670 by the WGCNA method.
doi:10.1371/journal.pone.0092395.g008

consistency and functional similarity were retained for building the high confident network, which was capable of avoiding biased results merely depending on expression. (iii) The cancer risks of modules were evaluated by using the proportion of significant tumor samples, which could be a new method to evaluate disease modules. The genes in cancer-risk modules could be potential disease genes, and might act as drug targets for the treatment of aggressive cancers. All genes of m46 were related with lung cancer. For instance, MCM7 is a significant subunit of MCM complex, which could be a novel therapeutic target in lung cancer [58]. Another gene BARD1, whose isoforms may be related to tumor initiation and invasive progression, was a more suitable neoteric prognostic marker for non-small-cell lung cancer [59]. KIF4A might hold a promise for the development of anticancer drugs and cancer vaccines as well as a prognostic biomarker in clinic [60]. For the genes in module m63, A2M was in limited and extended lung cancer patients compared to a nonsmoker and smoker control population [61], FABP4 was down-regulated in lung adenocarcinoma [62], and CASP1 affected the single-nucleotide polymorphisms, increasing the cancers risk [63]. (iv) The evaluation of samples' module-based disease risks is accuracy and robustness. Because our method integrated the differentially expressed genes, a co-expression network and functional similarities, the cancer-risk modules were closely related to the pathogenesis of cancer in the aspects of functions and interactions. On the functional level, the cancer-risk modules could reflect the functional classes related to diseases; on the interactional level, the cancer-risk modules could be very high correlated with the disease genes.

Additionally, we investigated the overlap between the cancer-risk modules and the protein compounds (Figure 9). The results of hypergeometric distribution analysis showed that 17 modules had significant overlap with 150 complexes ($p<0.05$). For example, module m46 shared genes with 24 complexes, among which 19 complexes had an overlap rate higher than 20%. The complex BRCA1_A recruited BRCA1 to DNA damage sites [64]. Partial depletion of Mcm proteins which were typically loaded in excessive number of locations led to cancers and stem cell deficiencies [65]. The expression of ubiquitin E3 ligase was associated with estrogen receptor (ER)-positive status in human breast tumors [66] (More modules and complex information are in the table S2). Our method will be more comprehensive considering protein-protein information to construct an integrated network and developing a module mining algorithm in the future.

In conclusion, this study presented a novel method to evaluate disease risks of samples based on cancer-risk modules and to analyze the relationships between the disease and modules. This method could provide assistance to the diagnosis

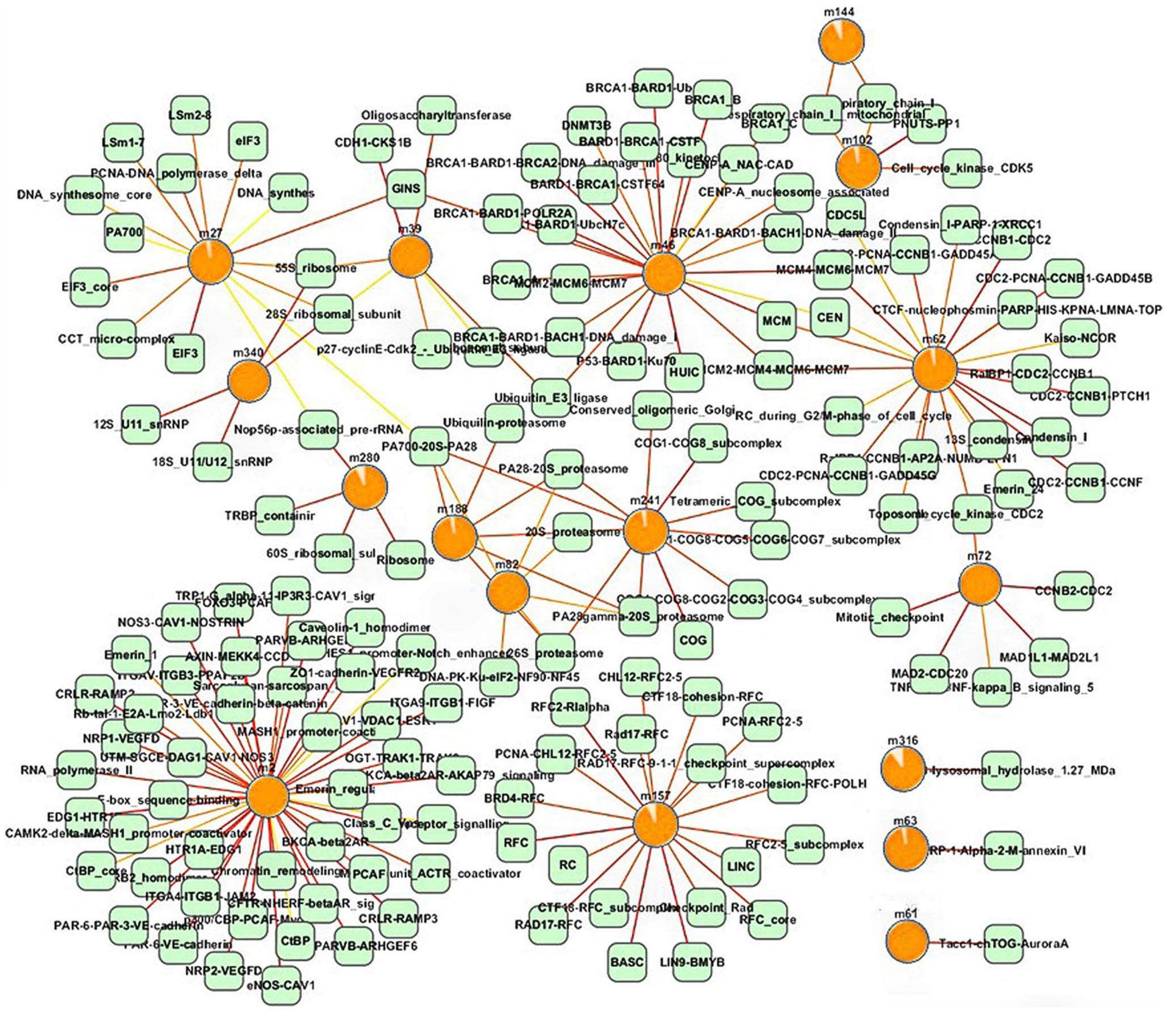


Figure 9. Overlapping relationship network of cancer-risk modules and complexes. The circles indicate cancer-risk modules, and the proportion of orange parts indicates cancer risk (M_{risk}). The green squares indicate complexes. Edges indicate cancer-risk modules and complexes sharing at least one gene. The more the number of shared genes are, the redder the edges are.

doi:10.1371/journal.pone.0092395.g009

and treatment of cancers and a new clue for revealing the cancer mechanisms.

Supporting Information

Table S1 The GO information of cancer-risk modules.
(DOC)

Table S2 Cancer-risk modules and complexs.
(DOC)

Table S3 The cancer-risk modules in the other four cancers.
(DOC)

Author Contributions

Conceived and designed the experiments: KL LZ LC. Performed the experiments: XJ ZM WL YH. Analyzed the data: XJ ZM WL MH YX. Contributed reagents/materials/analysis tools: XJ ZM WL CF XB LW YD DH. Wrote the paper: XJ ZM WL LC.

References

1. Zochbauer-Muller S, Fong KM, Virmani AK, Geradts J, Gazdar AF, et al. (2001) Aberrant promoter methylation of multiple genes in non-small cell lung cancers. Cancer Res 61: 249–255.
2. Compagno M, Lim WK, Grunn A, Nandula SV, Brahmachary M, et al. (2009) Mutations of multiple genes cause deregulation of NF-kappaB in diffuse large B-cell lymphoma. Nature 459: 717–721.
3. Santarius T, Shipley J, Brewer D, Stratton MR, Cooper CS (2010) A census of amplified and overexpressed human cancer genes. Nat Rev Cancer 10: 59–64.
4. Okabe H, Satoh S, Kato T, Kitahara O, Yanagawa R, et al. (2001) Genome-wide analysis of gene expression in human hepatocellular carcinomas using cDNA microarray: identification of genes involved in viral carcinogenesis and tumor progression. Cancer Res 61: 2129–2137.

5. Kettunen E, Anttila S, Seppanen JK, Karjalainen A, Edgren H, et al. (2004) Differentially expressed genes in nonsmall cell lung cancer: expression profiling of cancer-related genes in squamous cell lung cancer. Cancer Genet Cytogenet 149: 98–106.
6. Kreiman G (2004) Identification of sparsely distributed clusters of cis-regulatory elements in sets of co-expressed genes. Nucleic Acids Res 32: 2889–2900.
7. Li Y, St John MA, Zhou X, Kim Y, Sinha U, et al. (2004) Salivary transcriptome diagnostics for oral cancer detection. Clin Cancer Res 10: 8442–8450.
8. Ogata Y, Suzuki H, Sakurai N, Shibata D (2010) CoP: a database for characterizing co-expressed gene modules with biological information in plants. Bioinformatics 26: 1267–1268.
9. Shakhnovich BE, Reddy TE, Galinsky K, Mellor J, Delisi C (2004) Comparisons of predicted genetic modules: identification of co-expressed genes through module gene flow. Genome Inform 15: 221–228.
10. Dinu I, Potter JD, Mueller T, Liu Q, Adewale AJ, et al. (2009) Gene-set analysis and reduction. Brief Bioinform 10: 24–34.
11. Ein-Dor L, Zuk O, Domany E (2006) Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc Natl Acad Sci U S A 103: 5923–5928.
12. West M, Ginsburg GS, Huang AT, Nevins JR (2006) Embracing the complexity of genomic data for personalized medicine. Genome Res 16: 559–566.
13. Pittman J, Huang E, Dressman H, Horng CF, Cheng SH, et al. (2004) Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. Proc Natl Acad Sci U S A 101: 8431–8436.
14. Segal E, Friedman N, Koller D, Regev A (2004) A module map showing conditional activity of expression modules in cancer. Nat Genet 36: 1090–1098.
15. Hollen PJ, Gralla RJ (1996) Comparison of instruments for measuring quality of life in patients with lung cancer. Semin Oncol 23: 31–40.
16. Segal E, Friedman N, Kaminski N, Regev A, Koller D (2005) From signatures to models: understanding cancer using microarrays. Nat Genet 37 Suppl: S38–45.
17. Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, et al. (1993) The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 85: 365–376.
18. Desmedt C, Haibe-Kains B, Wirapati P, Buyse M, Larsimont D, et al. (2008) Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. Clin Cancer Res 14: 5158–5165.
19. Cho SB, Kim J, Kim JH (2009) Identifying set-wise differential co-expression in gene expression microarray data. BMC Bioinformatics 10: 109.
20. Watson M (2006) CoXpress: differential co-expression in gene expression data. BMC Bioinformatics 7: 509.
21. Fang G, Kuang R, Pandey G, Steinbach M, Myers CL, et al. (2010) Subspace differential coexpression analysis: problem definition and a general approach. Pac Symp Biocomput: 145–156.
22. Tesson BM, Breitling R, Jansen RC (2010) DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. BMC Bioinformatics 11: 497.
23. Zhang B, Horvath S (2005) A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 4: Article17.
24. Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9: 559.
25. Min JL, Nicholson G, Halgrimsdottir I, Almstrup K, Petri A, et al. (2012) Coexpression network analysis in abdominal and gluteal adipose tissue reveals regulatory genetic loci for metabolic syndrome and related phenotypes. PLoS Genet 8: e1002505.
26. de Jong S, Boks MP, Fuller TF, Strengman E, Janson E, et al. (2012) A gene co-expression network in whole blood of schizophrenia patients is independent of antipsychotic-use and enriched for brain-expressed genes. PLoS One 7: e39498.
27. Dewey FE, Perez MV, Wheeler MT, Watt C, Spin J, et al. (2011) Gene coexpression network topology of cardiac development, hypertrophy, and failure. Circ Cardiovasc Genet 4: 26–35.
28. Aoki K, Ogata Y, Shibata D (2007) Approaches for extracting practical information from gene co-expression networks in plant biology. Plant Cell Physiol 48: 381–390.
29. Carter SL, Brechbuhler CM, Griffin M, Bond AT (2004) Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics 20: 2242–2250.
30. Carlson MR, Zhang B, Fang Z, Mischel PS, Horvath S, et al. (2006) Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. BMC Genomics 7: 40.
31. Purmann A, Toedling J, Schueler M, Carninci P, Lehrach H, et al. (2007) Genomic organization of transcriptomes in mammals: Coregulation and cofunctionality. Genomics 89: 580–587.
32. Michalak P (2008) Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. Genomics 91: 243–248.
33. Ivliev AE, t Hoen PA, Sergeeva MG (2010) Coexpression network analysis identifies transcriptional modules related to proastrocytic differentiation and sprouty signaling in glioma. Cancer Res 70: 10060–10070.
34. Hong S, Dong H, Jin L, Xiong M (2011) Gene co-expression network and functional module analysis of ovarian cancer. Int J Comput Biol Drug Des 4: 147–164.
35. Teschendorff AE, Gomez S, Arenas A, El-Ashry D, Schmidt M, et al. (2010) Improved prognostic classification of breast cancer defined by antagonistic activation patterns of immune response pathway modules. BMC Cancer 10: 604.
36. Singh D, Febbo PG, Ross K, Jackson DG, Manola J, et al. (2002) Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1: 203–209.
37. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, et al. (2009) NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res 37: D885–890.
38. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102: 15545–15550.
39. Reverter A, Chan EK (2008) Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. Bioinformatics 24: 2491–2497.
40. Shi Z, Derow CK, Zhang B (2010) Co-expression module analysis reveals biological processes, genomic gain, and regulatory mechanisms associated with breast cancer progression. BMC Syst Biol 4: 74.
41. Su LJ, Chang CW, Wu YC, Chen KC, Lin CJ, et al. (2007) Selection of DDX5 as a novel internal control for Q-RT-PCR from microarray data using a block bootstrap re-sampling scheme. BMC Genomics 8: 140.
42. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29.
43. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, et al. (2009) Human Protein Reference Database—2009 update. Nucleic Acids Res 37: D767–772.
44. Mao L, Van Hemert JL, Dash S, Dickerson JA (2009) Arabidopsis gene co-expression network and its functional modules. BMC Bioinformatics 10: 346.
45. Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF (2007) A new method to measure the semantic similarity of GO terms. Bioinformatics 23: 1274–1281.
46. Balding DJ (2006) A tutorial on statistical methods for population association studies. Nat Rev Genet 7: 781–791.
47. Zhang L, Li W, Song L, Chen L (2010) A towards-multidimensional screening approach to predict candidate genes of rheumatoid arthritis based on SNP, structural and functional annotations. BMC Med Genomics 3: 38.
48. Reimand J, Tooming L, Peterson H, Adler P, Vilo J (2008) GraphWeb: mining heterogeneous biological networks for gene modules with functional significance. Nucleic Acids Res 36: W452–459.
49. Samuel Lattimore B, van Dongen S, Crabbe MJ (2005) GeneMCL in microarray analysis. Comput Biol Chem 29: 354–359.
50. Zaykin DV (2011) Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. J Evol Biol 24: 1836–1841.
51. Huang da W, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37: 1–13.
52. Sunaga N, Shames DS, Girard L, Peyton M, Larsen JE, et al. (2011) Knockdown of oncogenic KRAS in non-small cell lung cancers suppresses tumor growth and sensitizes tumor cells to targeted therapy. Mol Cancer Ther 10: 336–346.
53. Roth JA, Nguyen D, Lawrence DD, Kemp BL, Carrasco CH, et al. (1996) Retrovirus-mediated wild-type p53 gene transfer to tumors of patients with lung cancer. Nat Med 2: 985–991.
54. Kendall RL, Rutledge RZ, Mao X, Tebben AJ, Hungate RW, et al. (1999) Vascular endothelial growth factor receptor KDR tyrosine kinase activity is increased by autophosphorylation of two activation loop tyrosine residues. J Biol Chem 274: 6453–6460.
55. Massarelli E, Varella-Garcia M, Tang X, Xavier AC, Ozburn NC, et al. (2007) KRAS mutation is an important predictor of resistance to therapy with epidermal growth factor receptor tyrosine kinase inhibitors in non-small-cell lung cancer. Clin Cancer Res 13: 2890–2896.
56. Solis LM, Behrens C, Dong W, Suraokar M, Ozburn NC, et al. (2010) Nrf2 and Keap1 abnormalities in non-small cell lung carcinoma and association with clinicopathologic features. Clin Cancer Res 16: 3743–3753.
57. Perner S, Wagner PL, Demichelis F, Mehra R, Lafargue CJ, et al. (2008) EML4-ALK fusion lung cancer: a rare acquired event. Neoplasia 10: 298–302.
58. Croce CM (2009) Causes and consequences of microRNA dysregulation in cancer. Nat Rev Genet 10: 704–714.
59. Zhang YQ, Bianco A, Malkinson AM, Leoni VP, Frau G, et al. (2012) BARD1: an independent predictor of survival in non-small cell lung cancer. Int J Cancer 131: 83–94.
60. Taniwaki M, Takano A, Ishikawa N, Yasui W, Inai K, et al. (2007) Activation of KIF4A as a prognostic biomarker and therapeutic target for lung cancer. Clin Cancer Res 13: 6624–6631.
61. Marchandise FX, Mathieu B, Francis C, Sibille Y (1989) Local increase of antiprotease and neutrophil elastase-alpha 1-proteinase inhibitor complexes in lung cancer. Eur Respir J 2: 623–629.
62. Wang G, Ye Y, Zheng W, Ma W (2010) [Identification of candidate genes for lung adenocarcinoma using Toppgene]. Zhongguo Fei Ai Za Zhi 13: 282–286.
63. Dong LM, Brennan P, Karami S, Hung RJ, Menashe I, et al. (2009) An analysis of growth, differentiation and apoptosis genes with risk of renal cancer. PLoS One 4: e4895.
64. Wang B, Hurov K, Hofmann K, Elledge SJ (2009) NBA1, a new player in the Brca1 A complex, is required for DNA damage resistance and checkpoint control. Genes Dev 23: 729–739.

65. Rusiniak ME, Kunnev D, Freeland A, Cady GK, Pruitt SC (2012) Mcm2 deficiency results in short deletions allowing high resolution identification of genes contributing to lymphoblastic lymphoma. Oncogene 31: 4034–4044.

66. Kona FR, Stark K, Bisoski L, Buac D, Cui Q, et al. (2012) Transcriptional activation of breast cancer-associated gene 2 by estrogen receptor. Breast Cancer Res Treat 135: 495–503.