Hans Journal of Computational Biology 计算生物学, 2016, 6(2), 33-40
Published Online June 2016 in Hans. http://www.hanspub.org/journal/hjcb
http://dx.doi.org/10.12677/hjcb.2016.62005

Hans汉斯

The Construction and Analysis of Gene Co-Expression Network in Lung Cancer

Yuanyuan Zhai, Yingli Chen*, Jixian Xue

School of Physical Science and Technology, Inner Mongolia University, Hohhot Inner Mongolia
Email: *stchenyl@imu.edu.cn

Received: Jun. 6th, 2016; accepted: Jun. 20th, 2016; published: Jun. 28th, 2016

Abstract

Lung cancer, a complex molecular network disease, is a malignant tumor with the highest incidence and mortality around the world at present. In order to further understand the pathogenic molecular mechanism of lung cancer, we firstly identified differentially expressed gene (DEG) between cancer tissue and the corresponding adjacent normal tissue. Then, we used weighted gene co-expression network analysis (WGCNA) to screen for the DEG. In total, eight gene modules of DEG were detected and hub genes were identified. By calculating the Pearson's correlation coefficient between module eigengene and sample traits, we obtained the blue module which was highly associated with lung cancer, and found the hub gene of blue module was carbonic anhydrase 4. Hub gene plays an important role in the blue module. By using Database for Annotation, Visualization and Integrated Discovery (DAVID), the Gene Ontology (GO) enrichment analysis and KEGG pathway analysis were performed for blue module. The analysis of GO showed that blue module played important roles in biological functions, such as regulation of Rho protein signal transduction, biological adhesion, and carbohydrate binding. The analysis of KEGG indicated that blue module took part in the pathways of axon guidance and O-Glycan biosynthesis. The results showed that high correlations modules of lung cancer and hub gene identified in this paper played a potentially important role in the development of lung cancer.

Keywords

Lung Cancer, Gene Co-Expression Network, Hub Gene, WGCNA

*通讯作者。

文章引用: 翟媛媛, 陈颖丽, 薛济先. 与肺癌相关的基因共表达网络的构建与分析[J]. 计算生物学, 2016, 6(2): 33-40.
http://dx.doi.org/10.12677/hjcb.2016.62005

与肺癌相关的基因共表达网络的构建与分析

翟媛媛，陈颖丽*，薛济先

内蒙古大学物理科学与技术学院，内蒙古 呼和浩特
Email: *stchenyl@imu.edu.cn

收稿日期：2016年6月6日；录用日期：2016年6月20日；发布日期：2016年6月28日

摘 要

肺癌是目前世界范围内发病率和死亡率最高的恶性肿瘤，它是一种复杂的分子网络疾病。为了进一步了解肺癌致病的分子机制，我们使用加权基因共表达网络分析(WGCNA)方法，对肺癌组织与其癌旁正常组织的差异表达基因进行分析，进而对差异表达基因进行模块的划分以及枢纽基因(hub gene)的识别，共得到了八个模块。通过计算每个模块特征向量基因(module eigengene)与样本特征的皮尔森相关系数，最终得到一个与肺癌高关联的模块(blue模块)，发现blue模块的枢纽基因为碳酸酐酶4 (carbonic anhydrase 4, CA4)，这一枢纽基因在模块中起着重要的作用。使用在线工具DAVID (Database for Annotation, Visualization and Integrated Discovery)对blue模块进行GO功能富集及KEGG通路分析。GO分析显示blue模块具有Rho 蛋白的信号转导调控、生物粘附、糖结合等生物功能；KEGG分析显示blue模块参与了轴突导向和O 型聚糖的生物合成通路。这些分析结果表明，文中识别的肺癌高关联模块和枢纽基因在肺癌的发生发展过程中起着潜在的重要作用。

关键词

肺癌，基因共表达网络，枢纽基因，WGCNA

1. 引言

肺癌是目前全世界最常见的恶性肿瘤之一[1]，在男性中其发病率居所有恶性肿瘤首位，也是导致男性恶性肿瘤死亡的首要原因，在女性中发病率居第 4，死亡率居第 2，并且 5 年生存率很低[2]。目前有效的治疗是全肺切除加适当的化疗策略。因而对于肺癌的治疗最重要的仍然是寻找有效的早期诊断和指导预后的标志物，为对抗癌细胞的侵袭和转移提供理论依据。基因共表达网络分析已经成功的用于识别参与多种疾病的重要基因、生物学过程以及通路。加权基因共表达网络分析(WGCNA)是构建基因共表达网络的常用方法，它是一种强大的基于系统生物学思想的方法，用于解析分子作用机制和网络关系，并将高度关联的基因聚类到一个模块(module) [3]。WGCNA 方法自诞生以来已被成功应用于多种生物学问题的研究，并取得了许多重要发现[4]。例如，Liao 等[5]利用 WGCNA 构建了编码－非编码基因共表达网络。Plaisier 等[6]通过 WGCNA 方法，并结合遗传学标记数据识别出了 USF1 和 FADS3 是家族性复合高脂血症的相关通路基因。此外，使用 WGCNA 计算基因之间相关性时，假阳性率很低。因此，它已经普遍的被应用于复杂疾病的研究，例如，子宫内膜癌、乳腺癌、精神分裂症、食管鳞癌等[7] [8]。

在这篇文章中，我们使用 WGCNA 来研究肺癌基因共表达谱与样本特征之间的相关性，得到一个与肺癌高关联的模块，对这个模块进行 GO 功能富集分析和 KEGG 通路分析，并找到这个模块的枢纽基因，

进而对枢纽基因的功能进行分析。结果表明，枢纽基因可能作为肺癌有效的早期诊断分子和指导预后的标志物。

2. 数据和方法

2.1. 数据

本文数据集源自2010年Tzu-Pin Lu等的工作[9]。该数据来自NCBI中的GEO数据库(Gene Expression Omnibus，http://www.ncbi.nlm.nih.gov/geo)。GEO数据库包括原始数据和预处理后的数据，本文使用的数据是预处理后的数据。数据平台编号是GPL570，选取的数据集编号是GSE19804，芯片类型是HG_U133_Plus_2。该数据集中包含120个样本，其中60个是肺癌组织样本，而另60个是相应的癌旁正常组织样本，这些样本均来自于台湾无抽烟史的女性肺癌患者。在本文中，我们使用R语言GSEquery和Limma软件包对下载的数据进行处理，通过设定相关阈值得到差异表达基因。

2.2. 加权基因共表达网络的构建

WGCNA是使用基因表达数据来构建无尺度网络的系统生物学方法。其基本思路如下[3] [10]。首先，构建基因表达相似性矩阵，即计算两两基因之间皮尔森相关系数的绝对值，使用公式1计算基因i和基因j之间的皮尔森相关系数，其中i和j分别是第i个基因和第j个基因的表达量。

$$S_{ij} = \left| \frac{1 + cor\left(x_i + y_j\right)}{2} \right| \tag{1}$$

然后，使用公式2将基因表达相似性矩阵转换成邻接矩阵，网络类型为signed，其中β为软阈值，其实就是将每对基因的皮尔森相关系数β次方。这一步能够从指数级别强化强相关性和减弱弱相关性。

$$a_{ij} = \left| \frac{1 + cor\left(x_i + y_j\right)}{2} \right|^{\beta} \tag{2}$$

下一步使用公式3将邻接矩阵转换成拓扑矩阵，拓扑重叠(topological overlap measure, TOM)用来描述基因之间的关联程度。

$$TOM = \frac{\sum_{u \neq ij} a_{iu} a_{uj} + a_{ij}}{\min\left(\sum_u a_{iu} + \sum_u a_{ju}\right) + 1 - a_{ij}} \tag{3}$$

1-TOM表示基因i和基因j之间的相异程度。使用1-TOM作为距离对基因进行层次聚类，然后使用动态剪切树的方法进行模块的识别。每个模块中最具有代表性的基因称为特征向量基因，简称ME，它代表了该模块内基因表达的整体水平，它是每个模块中的第一主成分，使用公式4来计算ME，其中i表示模块q中的基因，l表示模块q中的芯片样本。

$$ME = princomp\left(x_{il}^{(q)}\right) \tag{4}$$

我们用某个基因在所有样本中的表达谱与某个特征向量基因ME表达谱的皮尔森相关性来衡量这个基因在该模块中的身份，即模块身份(module membership)，简称MM。使用公式5计算MM，其中x_i表示第i个基因的表达谱，ME^q表示模块q的特征向量基因(ME)，MM_i^q表示了基因i在模块q中的身份，当$MM_i^q = 0$，则说明基因i不在模块q中，MM_i^q越接近+1或−1，则说明基因i与模块q高度相关。正负号表示了基因i与模块q是正相关还是负相关。

$$MM_i^q = cor\left(x_i, ME^q\right) \tag{5}$$

基因显著性(gene significance)，简称 GS，用来衡量基因与外部信息的关联程度，GS 越高表示基因越具有生物学意义，GS = 0，说明这个基因不参与所研究的生物学问题。

2.3. 关键共表达基因模块和枢纽基因的筛选

计算每个模块的特征向量基因(ME)与样本特征信息的皮尔森相关系数来确定关键模块。枢纽基因具有很高的连接度，它是一个模块中具有高连接度的一系列基因。WGCNA 的一个目的就是找出感兴趣模块的枢纽基因。一般来说，相对于全局网络，子模块的枢纽基因更具有生物学意义。已经证明模块身份(MM)可以用来衡量一个基因在某个模块中的重要性，并且 MM 与模块的连接度有着正相关的联系[11]。因此，我们根据 MM 值来选取枢纽基因，若某个基因在一个特定的模块中具有最大的|MM|值，则把这个基因当做枢纽基因。

2.4. GO 和 KEGG 通路富集分析

使用 DAVID (https://david.ncifcrf.gov/)对模块进行 GO 富集分析和 KEGG 通路分析，DAVID 是一个生物信息数据库，整合了生物学数据和分析工具，为大规模的基因或蛋白质列表提供系统综合的生物功能注释信息，帮助用户从中提取生物学信息。

3. 结果

3.1. 数据处理和差异表达基因的筛选

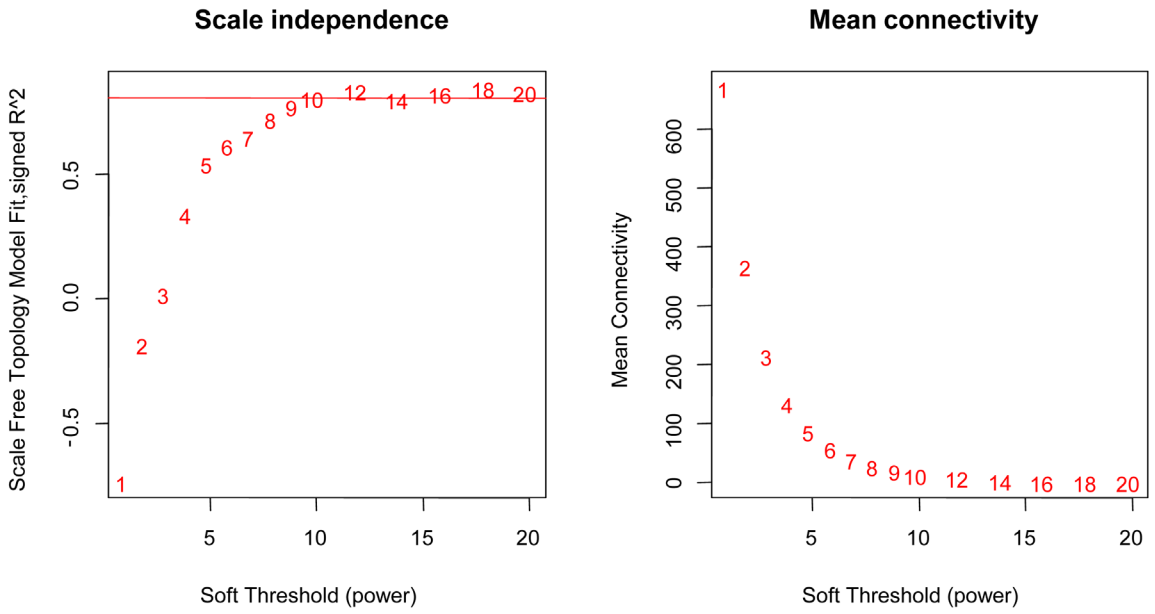
对于多个探针对应一个基因的情况，我们取平均值作为这个基因的表达值。首先将预处理后的数据转化为基因表达矩阵，其中行表示基因，列表示基因在不同样本中的表达值。再将样本构建成一个 120 维的样本特征向量，其中正常样本为“1”，癌症样本为“2”，基于样本特征向量，使用线性模型对基因表达矩阵中各行的基因表达向量进行拟合，再对拟合结果进行贝叶斯检验[12]。设阈值为|lgFC| > 1，共得到 1359 个差异表达基因。可使用 R 语言中的 Limma 软件包完成。

3.2. 加权基因共表达网络的构建

使用肺癌组织与相应的癌旁正常组织中 1359 个差异表达基因进行网络的构建，使用的方法为 R 语言中的 WGCNA 软件包。研究表明共表达网络符合无尺度网络，即出现连接度为 k 的节点的对数 log(k)与该节点出现的概率的对数 log(P(k))要负相关，且相关系数要大于 0.8。为了确保网络为无尺度网络，我们选择 $\beta = 12$ (图 1)。下一步将表达矩阵转换成邻接矩阵，然后再将邻接矩阵转换成拓扑矩阵，基于 TOM，我们使用 average-linkage 层次聚类法对基因进行聚类，使用动态切割算法(deep split = 4)从系统聚类树中识别模块，共得到了 9 个模块(图 2)，需要指出，grey 模块是无法聚集到其它模块的基因集合。图 3 验证了当 $\beta = 12$ 时，网络是否具有无尺度特性。图 3(a)显示了基因连接度的频率分布特征，图 3(b)显示了 log(k)和 log(p(k))成负相关，且相关系数大于 0.8，证实了所构建网络无尺度的特性。

3.3. 关键共表达基因模块和枢纽基因的筛选

计算每个模块的 ME 与样本特征的皮尔森相关系数，相关系数越高说明这个模块越重要(图 4)。图 4 中行表示每个模块的特征向量基因，列表示样本特征信息，其中第一列 tissue 表示样本组织来源，第二列 age 表示年龄，第三列 stage 表示肿瘤分期。从红色到绿色表示相关系数从高到低依次递减。每个小格子里的数字表示基因模块与相应特征的相关系数，括号中的数字表示 P 值。从图中我们可以得出 blue 模



(a) (b)

Figure 1. Analysis of network topology for various soft-thresholding powers
图 1. 分析不同阈值下网络的拓扑结构

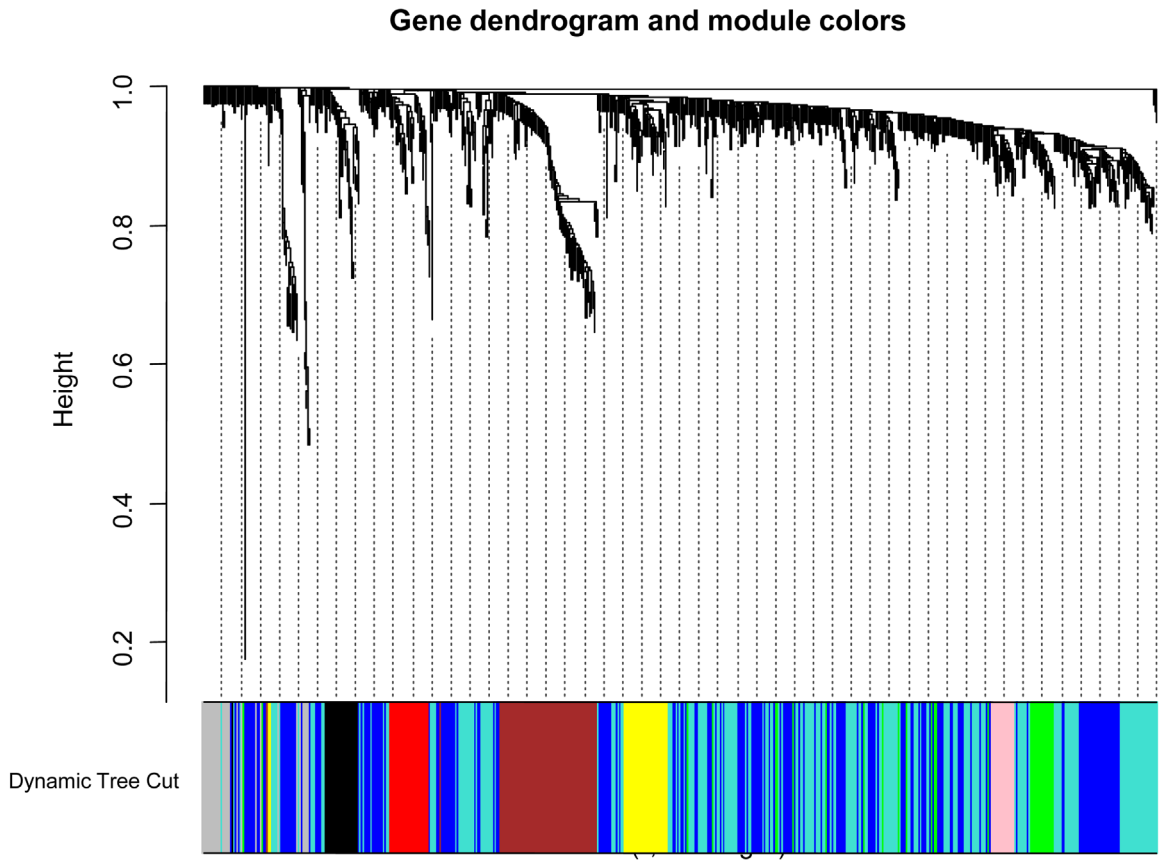
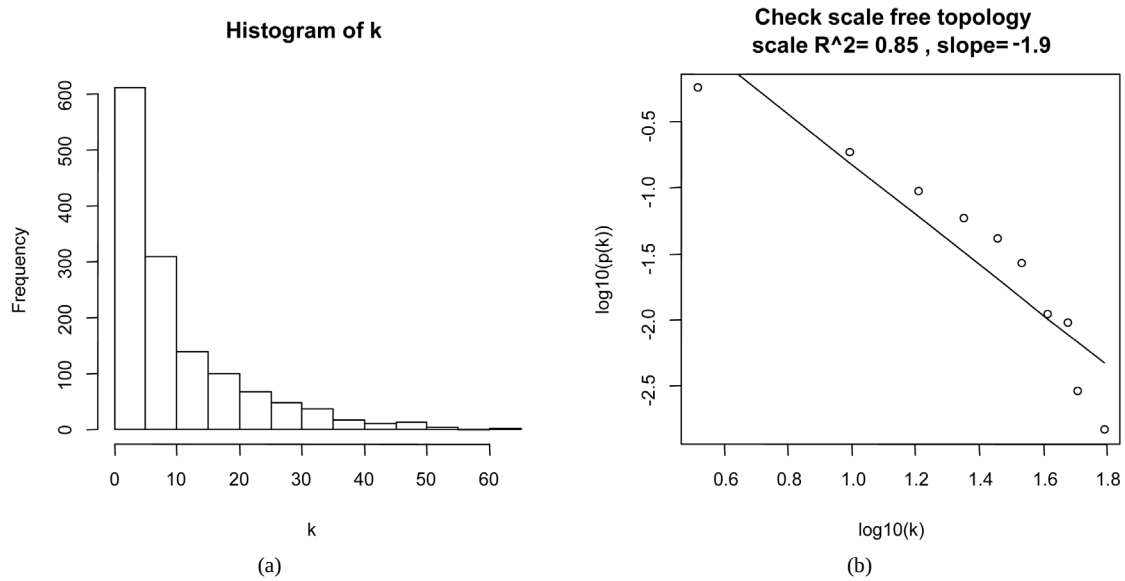


Figure 2. Gene dendrogram and module colors
图 2. 基因树状图以及模块颜色



(a) (b)

Figure 3. The test of property of scale-free network
图 3. 无尺度网络特性检验

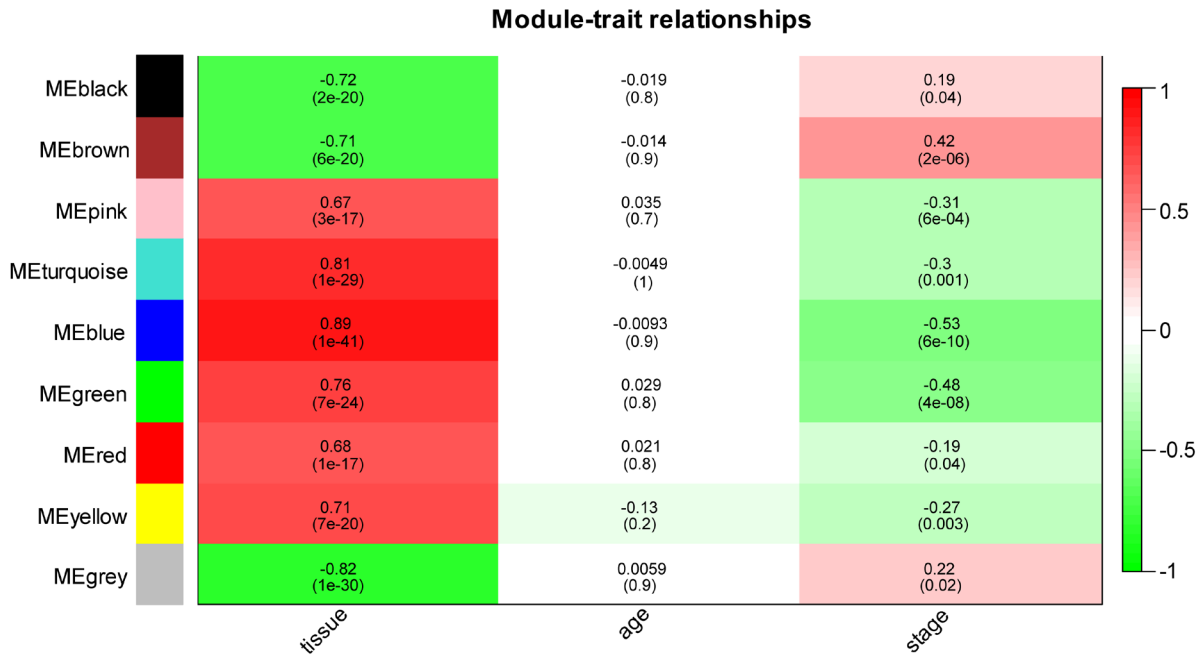


Figure 4. Module-trait associations
图 4. 模块-特征的相关性

块与样本组织来源(即是样本是否换肺癌)相关性最强，即 blue 模块为肺癌高关联模块，所以接下来我们主要针对这个模块进行后续的分析。计算所有基因与 blue 模块特征向量基因的相关性，可以得到 blue 模块的枢纽基因为碳酸酐酶 4 (MM^{blue}_{CA4} = 0.952)。

3.4. GO 功能富集分析和 KEGG 通路分析

GO 功能富集分析用于识别感兴趣模块所富集到的重要 GO 语义[13]。使用 DAVID 对 blue 模块进行

GO 富集分析以及 KEGG 通路分析，发现这个模块富集到一些重要的 GO 语义和生物学通路。GO 富集分析的结果见表 1，blue 模块富集到生物学过程(biological process)的语义为：Rho 蛋白的信号转导调控(regulation of Rho protein signal transduction)、生物粘附(biological adhesion)；富集到细胞组分(cellular component)的语义为：固有膜(intrinsic to membrane)、细胞外区域(extracellular region part)；富集到分子功能(molecular function)的语义为：碳水化合物的结合(carbohydrate binding)、Rho 鸟苷酸转换因子活性(Rho guanyl-nucleotide exchange factor activity)。KEGG 通路富集分析结果如见表 2，blue 模块富集到的通路为：轴突导向(axon guidance)、O 型聚糖的生物合成(O-Glycan biosynthesis)。

4. 讨论

在以上研究中，我们发现 blue 模块是肺癌高关联模块，并且找到了这个模块的枢纽基因。GO 富集分析结果显示，blue 模块显著富集的生物学过程是：Rho 蛋白的信号转导调控、生物粘附。KEGG 通路分析显著富集在：轴突导向、O 型聚糖的生物合成。粘附分子是指由细胞产生、存在于细胞表面、介导细胞与细胞间或细胞与基质间相互接触和结合的一类分子，大多为糖蛋白，少数为糖脂，分布于细胞表面或细胞外基质中。而 CA4 是 blue 模块的枢纽基因，CA4 基因编码产物为糖基膜锚定酶，是一种广泛存在于生物体中的一种酶，功能是催化碳水化合物可逆的分解为碳酸氢盐和氢离子，主要分布在肺毛细血管腔表面和肾小管表面。许多研究表明，碳酸酐酶 4 参与肿瘤发生、转移，并与实体瘤的不良预后相关。例如，在实体瘤的一期临床阶段，使用磺酰胺抑制剂抑制 CA4 表达，可以调控肿瘤的转移[14]。KEGG 通路分析主要富集在轴突导向，已有一些研究表明：轴突导向分子，特别是 semaphorin 蛋白与肿瘤的发生、转移以及细胞的凋亡密切相关[15]。Semaphorin 蛋白是一种分泌或跨膜糖基蛋白，可影响肿瘤细胞的迁移和增殖以及影响肿瘤血管生成和淋巴细胞趋化性，并且可作为肺癌发生的抑制子。CA4 是人类 12 种活性同工酶中的一种，1/4 的 CA4 存在于某些内皮和外皮细胞的细胞外表面，它是唯一一个不是通过跨膜结构域而是通过糖基锚定在质膜上的碳酸酐酶[16]，正好参与以上富集到生物粘附过程和轴突导向通路。这些结论表明，文中识别的肺癌高关联模块以及这个模块的枢纽基因在肺癌患者的发生过程中起着重要的作用。

Table 1. Most enriched functional annotations of the blue module
表 1. Blue 模块显著富集的 GO 语义

Category	Term name	P-value
GOTERM_BP_FAT	regulation of Rho protein signal transduction	3.8E−5
GOTERM_BP_FAT	biological adhesion	3.0E−4
GOTERM_CC_FAT	intrinsic to membrane	4.6E−6
GOTERM_CC_FAT	extracellular region part	9.9E−6
GOTERM_FM_FAT	carbohydrate binding	5.4E−4
GOTERM_FM_FAT	Rho guanyl-nucleotide exchange factor activity	9.3E−4

Table 2. Most enriched pathway of the blue module
表 2. Blue 模块显著富集的通路

Category	Term	P-value
KEGG_PATHWAY	axon guidance	2.7E−4
KEGG_PATHWAY	O-Glycan biosynthesis	6.9E−4

致　谢

感谢国家自然科学基金(61361015)和教育部第 46 批留学回国人员科研启动基金的支持。

参考文献 (References)

[1] Torre, L.A., Bray, F., Siegel, R.L., Ferlay, J., Lortet-Tieulent, J. and Jemal, A. (2015) Global Cancer Statistics, 2012. *CA-A Cancer Journal for Clinicians*, **65**, 87-108. http://dx.doi.org/10.3322/caac.21262

[2] Jemal, A., Bray, F., Center, M.M., Ferlay, J., Ward, E. and Forman, D. (2011) Global Cancer Statistics. *CA-A Cancer Journal for Clinicians*, **61**, 69-90. http://dx.doi.org/10.3322/caac.20107

[3] Langfelder, P. and Horvath, S. (2008) WGCNA: An R Package for Weighted Correlation Network Analysis. *BMC Bioinformatics*, **9**, 559-572. http://dx.doi.org/10.1186/1471-2105-9-559

[4] 钟诗龙, 伍虹, 杨敏, 等. 用权重基因共表达网络分析识别心脏重构关键节点基因[J]. 中国药理学通报, 2011, 27(10): 1358-1362.

[5] Lin, R.C., Weeks, K.L., Kang, S., Miao, R., Xiao, H., Zhao, G., Luo, H., Du, D., Zhao, H., *et al*. (2011) Large-Scale Prediction of Long Non-Coding RNA Functions in a Coding-Non-Coding Gene Co-Expression Network. *Nucleic Acids Research*, **39**, 3864-3878. http://dx.doi.org/10.1093/nar/gkq1348

[6] Plaisier, C.L., Horvath, S., Huertas-Vazquez, A., *et al*. (2009) A Systems Genetics Approach Implicates USF1, FADS3, and Other Causal Candidate Genes for Familial Combined Hyperlipidemia. *PLoS Genets*, **5**, e1000642. http://dx.doi.org/10.1371/journal.pgen.1000642

[7] Xing, Y., Zhang, J., Lu, L., Li, D., Wang, Y., Huang, S., *et al*. (2016) Identification of Hub Genes of Pneumocyte Senescence Induced by Thoracic Irradiation Using Weighted Gene Coexpression Network Analysis. *Molecular Medicine Reports*, **13**, 107-116.

[8] 王攀. 加权基因共表达网络(WGCNA)在食管鳞癌中的应用[D]: [博士学位论文]. 北京: 北京协和医院, 2014.

[9] Lu, T.P., Tsai, M.H., Lee, J.M., Hsu, C.P., Chen, P.C., Lin, C.W., *et al*. (2010) Identification of a Novel Biomarker, SEMA5A, for Non-Small Cell Lung Carcinoma in Nonsmoking Women. *Cancer Epidemiology Biomarkers & Prevention*, **19**, 2590-2597. http://dx.doi.org/10.1158/1055-9965.EPI-10-0332

[10] Zhang, B. and Horvath, S. (2005) A General Framework for Weighted Gene Co-Expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology*, **4**, 1-45. http://dx.doi.org/10.2202/1544-6115.1128

[11] Xing, Y.X., Zhang, J.L., Lu, L., Li, D.G., *et al*. (2016) Identification of Hub Genes of Pneumocyte Senescence Induced by Thoracic Irradiation Using Weighted Gene Coexpression Network Analysis. *Molecular Medicine Reports*, **13**, 107-116.

[12] 梁栋, 邢永强, 蔡禄. 肾肿瘤相关基因的共表达网络构建与分析[J]. 中国生物工程杂志, 2016, 36(2): 30-37.

[13] Lecca, P. and Re, A. (2015) Detecting Modules in Biological Networks by Edge Weight Clustering and Entropy Significance. *Front Genet*, **6**, 265-277. http://dx.doi.org/10.3389/fgene.2015.00265

[14] Supuran, C.T. (2016) How Many Carbonic Anhydrase Inhibition Mechanisms Exist. *Journal of Enzyme Inhibition and Medicinal Chemistry*, **31**, 345-360. http://dx.doi.org/10.3109/14756366.2015.1122001

[15] Chedotal, A., Kerjan, G. and Moreau-Fauvarque, C. (2005) The Brain within the Tumor: New Roles for Axon Guidance Molecules in Cancers. *Cell Death Differ*, **12**, 1044-1056. http://dx.doi.org/10.1038/sj.cdd.4401707

[16] Waheed, A. and Sly, W.S. (2014) Membrane Associated Carbonic Anhydrase IV (CA IV): A Personal and Historical Perspective. *Subcell Biochem*, **75**, 157-179. http://dx.doi.org/10.1007/978-94-007-7359-2_9