遗传 HEREDITAS (Beijing) 2013 年 3 月, 35(3): 333－342
ISSN 0253-9772 www.chinagene.cn

研究报告

DOI: 10.3724/SP.J.1005.2013.00333

基于双聚类挖掘癌症共享的基因功能模块

张凡[1], 林爱华[1], 林美华[2], 丁元林[2], 饶绍奇[1,2]

1. 中山大学公共卫生学院, 广州 510080;
2. 广东医学院公共卫生学院, 东莞 523808

摘要: 基因多效性是癌症遗传机制中的普遍现象, 但罕见系统性的分析。文章提出采用双聚类挖掘基因功能模块的新思路探索癌症的共享分子机制和不同癌症间的关系。获取 20 种癌症的基因表达数据, 应用改良 *t* 检验和倍数法筛选出至少在两种癌症中差异表达的基因, 得到 10417×20 的数据矩阵; 采用双聚类方法获得 22 个癌症共享的基因簇; 进一步富集分析得到 17 个基因功能模块(Bonferroni 校正后 $P<0.05$), 主要参与有丝分裂染色单体分离的调控、细胞分化、免疫和炎症反应、胶原纤维组织等生物过程; 主要执行 ATP 结合和微管活动、MHC Ⅱ类受体活性、肽链内切酶抑制活性等分子功能; 活动区域主要在细胞骨架、染色体、MHC Ⅱ蛋白质复合体、中间丝蛋白、胶原纤维等。基于模块构建癌症相关网络, 显示胃癌、卵巢腺癌、宫颈鳞癌和间皮瘤等之间相关程度较高, 而两种血液系统癌症(急性髓细胞性白血病与多发性骨髓瘤)分子机制与其他癌症存在较大差异。可见癌症共享的基因功能模块与多种生物机制有关, 癌症之间相似性可能与组织起源、共同的致癌机制等有关。文章提出的基因多效性分析方法有助于解释人类复杂性疾病的共享分子机制。

关键词: 癌症; 基因多效性; 双聚类; 基因功能模块

Identification of gene functional modules shared by cancers based on biclustering

ZHANG Fan[1], LIN Ai-Hua[1], LIN Mei-Hua[2], DING Yuan-Lin[2], RAO Shao-Qi[1,2]

1. *Shool of Public Health*, *Sun Yat-Sen University*, *Guangzhou* 510080, *China*;
2. *School of Public Health*, *Guangdong Medical College*, *Dongguan* 523808, *China*

Abstract: Pleiotropy is a common phenomenon in the genetics of cancers, which is rarely systematically evaluated. A novel idea for identifying shared gene functional modules using biclustering was proposed in this paper to explore the common molecular mechanisms among cancers and the relationships between different types of cancers. Gene expression datasets for 20 cancers were obtained. And genes differentially expressing in at least two types of cancers were selected using both moderated *t*-statistic and fold change to construct a 10417 × 20 matrix (gene-cancer matrix). 22 gene clusters shared by cancers were found by using the biclustering method. Further, Gene Ontology (GO)-based enrichment analysis

收稿日期: 2012–10–09; 修回日期: 2012–11–13
基金项目: 广东省科技计划攻关项目(编号: 2009A030301004), 东莞市科技重点项目(编号: 201108101015), 广东医学院基金项目(编号: XG1001, XZ1105, STIF201122)和国家自然科学基金项目(编号: 30830104, 31071166)资助
作者简介: 张凡, 硕士, 研究方向: 流行病与卫生统计学。Tel: 13760824092; E-mail: lemon_fan@163.com
通讯作者: 饶绍奇, 教授, 博士生导师, 研究方向: 遗传统计与生物信息学方向。E-mail: raoshaoq@gdmc.edu.cn
网络出版时间: 2013-1-8 9:36:09
URL: http://www.cnki.net/kcms/detail/11.1913.R.20130108.0936.002.html

identified 17 gene functional modules (Bonferroni corrected $P < 0.05$). The involved biological processes primarily included regulation of chromatids separation during mitosis, cell differentiation, immune and inflammatory response, and collagen fibril organization. These modules undertook molecular functions of ATP binding and microtubule motor activity, MHC class II receptor activity, endopeptidase inhibitor activity and so on. And their activity sites were mostly located in cytoskeleton, chromosome, MHC protein complex, intermediate filament, fibrillar collagen and so on. The network constructed based on these modules indicates that gastric cancer, ovarian adenocarcinoma, cervical cancer and mesothelioma were highly relevant to each other. However, the molecular mechanisms of two hematologic malignancies (acute myeloid leukemia and multiple myeloma) seem very different from other cancers. It can be seen that gene functional modules shared by cancers are associated with many biological mechanisms, and similarities among cancers are probably attributed to cellular origin and shared carcinogenic mechanisms. The proposed method for analysis of pleiotropy in this paper will help understand the common molecular mechanisms for complex human diseases.

Keywords: cancer; pleiotropy; biclustering; gene functional module

癌症(Cancer)是一类临床表型存在很大差异的恶性肿瘤的统称，主要表现为癌细胞的无限制增殖，被普遍认为是一种遗传病，体细胞基因组分子改变的累积是其发展的基础。癌症具有复杂的遗传机制，基于HuGE Navigator数据库[1]，截止到 2012 年 9 月共有 4 166 个基因被报道可能参与肿瘤的发生发展。尽管如此，癌症的确切分子机制和不同癌症类型之间的关系仍需进一步的研究。

基因多效性(Pleiotropy)是指单个基因的突变效应能够作用于两种或多种性状的生物学现象[2]，在癌症的遗传机制中普遍存在。综合大量连锁分析和关联分析的结果已经证实不同癌症表型可共享风险基因，即单个基因的突变可能引起不同的病理效应而参与多种癌症的发生发展[3]。例如抑癌基因*TP53*，可诱导细胞周期停滞、凋亡、衰老和DNA修复等生物过程，是已知的与多种癌症表型相关的多效性基因[4]。近年来，分子技术的发展积累了大量有关癌症的组学数据，为系统分析癌症和易感基因之间的关系提供可能。Goh等[5] 基于OMIM(Online Mendelian Inheritance in Man)数据库构建了人类疾病网络，指出同一类复杂性疾病，特别是癌症更倾向于共享分子机制，不同癌症表型之间通过常见癌基因或抑癌基因等而紧密相连，形成密切相关的一大类疾病；研究还指出大部分易感基因可能是多效性基因，而且随着高通量数据的进一步积累，多效性基因的数目将继续增长。由于多效性基因能够作用于两种或多种不同的癌症表型，产生相似甚至是相反效应，这对于指导和设计更为有效的早期预防、诊断和治疗癌症的方法和手段具有重要的意义。

以往有关癌症多效性基因的研究多局限于对现有知识库(如OMIM，NHGRI)的分析，受到已有知识的限制[5,6]；或者独立分析少数几种癌症表型的易感基因，无法系统地了解多种癌症的遗传关系[7,8]。目前越来越多研究指出，基因是通过模块，即一簇具有相关功能的基因的集合协同发挥功能，对模块的干扰可能会导致相似的多种疾病表型[9]。基于这样的假设，本文采用双聚类方法，综合分析了 20 种不同表型的癌症的DNA芯片数据集，寻找癌症共享的基因功能模块，以此探索癌症共享的分子机制和癌症之间的遗传关系。

1 数据与方法

1.1 数据来源

癌症基因表达谱数据来自GEO(http://www.ncbi.nlm.nih.gov/geo/)数据库，研究设计均为病例对照研究，共有 862 张芯片，来自Affymetrix的 3 种实验平台，包括 20 种常见的癌症表型(表 1)，组织来源均是相应癌组织，急性髓细胞性白血病和多发性骨髓瘤来自骨髓。本文下载原始的基因表达数据即CEL文件，统一用RMA(Robust Multi-array Analysis)方法对表达数据进行预处理，包括背景校正，标准化和PM(Perfect Match)探针值的校正[10]，在R“affy”软件包完成[11]。由于数据库只提供间皮瘤、皮肤鳞状细胞

表 1　20 种癌症的基因表达数据

癌症类型	#对照\病例	参考文献 (PMID)	实验平台 (Affymetrix)	数据来源 (GEO 登录号)
乳腺癌	12\54	22832278	HG-U133_Plus_2	GSE29431
卵巢腺癌	12\12	20040092	HG-U133_Plus_2	GSE14407
宫颈鳞癌	10\21	17974957	HG-U133A	GSE7803
肺腺癌	49\58	18297132	HG-U133A	GSE10072
鼻咽癌	10\31	16912175	HG-U133_Plus_2	GSE12452
间皮瘤	9\40	15920167	HG-U133A	GSE2549
胃癌	31\38	19081245	HG-U133_Plus_2	GSE13911
食管癌	5\5	—	HG-U133_Plus_2	GSE17351
大肠癌	8\15	19461970	HG-U133_Plus_2	GSE4183
肝细胞性肝癌	10\10	21747116	HG-U133_Plus_2	GSE29721
胰腺癌	7\25	22261810	HG-U133_Plus_2	GSE32676
肾透明细胞癌	23\32	16115910	HG-U133A	GSE15641
肾上腺皮质癌	10\33	22800756	HG-U133_Plus_2	GSE33371
膀胱癌	9\13	15173019	HG-U133A	GSE3167
皮肤鳞状细胞癌	6\5	20231500	HG-U133A	GSE2503
黑色素瘤	7\44	16243793	HG-U133A	GSE3189
脑胶质瘤	5\30	21406405	HG-U133_Plus_2	GSE15824
脂肪肉瘤	9\89	20601955	HG-U133A	GSE21122
急性髓细胞性白血病	38\26	17910043	HG-U133A	GSE9476
多发性骨髓瘤	5\6	22517906	HG-U133A_2	GSE24870

注：参考文献列出文献在 PubMed 数据库的索引号，其中食管癌表达数据的参考文献缺如。

癌和黑色素瘤预处理后数据，预处理方法参照相应文献。

1.2　基因初筛

联合应用改良的t检验(Moderated t-statistic)[12]和倍数法(Fold Change)评估基因在病例对照中的差异表达程度，对芯片上的基因进行初筛。为了控制多重检验造成的假阳性，t检验所得P值采用FDR (False discovery rate)法进行校正。倍数法用基因在病例的表达均数与正常对照的表达均数的比值的对数(Log2-fold Change, logFC)表示：

$$\log FC_i = \log_2\left(\frac{\bar{x}_i}{\bar{y}_i}\right)$$

其中 $\bar{x}_i$ 和 $\bar{y}_i$ 分别是基因 i 在病例和正常对照中的表达均数。本文将芯片上的探针映射到基因(采用相应实验平台的 Affymetrix 注释文件)，若多个探针对应 1 个基因，则用取 P 值较小的探针 logFC 值作为该基因的差异表达值；然后对不同芯片平台上的基因取交集，并删除满足下列任一条件的基因 1)校正后的P值在不多于一种癌症数据中小于 0.01；2)logFC的绝对值在不多于一种癌症数据中大于 0.585，即FC>1.5。以上分析由R “limma” 软件包完成[12]。

1.3　癌症共享基因簇的挖掘

癌症共享的基因簇定义为在多种癌症中差异表达的基因的集合。本文利用基因在 20 种癌症表型的相对表达值构建数据矩阵，其中第i行第j列的元素为第i个基因在第j种癌症表型的logFC值，然后应用双聚类算法，即 SAMBA(Statistical-Algorithmic Method for Bicluster Analysis)寻找多种癌症表型共享的基因簇。SAMBA算法与传统聚类相比，允许综合异类数据；并且基因或表型可以属于多个簇，簇之间可以存在重叠，更符合生物系统内部相关的特性[13]。

SAMBA算法的基本思想是结合统计模型将一个双聚类问题转化为在一个二分图中搜索稠密子图的问题[14]，主要包括 3 个阶段：(I)将数据矩阵(图 1A)转化为二部图G=(U,V,E)，其中U和V为两个不相交顶点的集合，分别表示癌症表型和基因，E是二部图

G中相连的边，表示基因v在癌症u中差异表达，即顶点对$(u,v)\in E$ (图 1B)，然后基于对数似然比(log likelihood ratio)计算G中每对顶点的权重，则子图$H=(U',V',E')$的得分为：

$$\log L(H)=\sum_{(u,v)\in E'}\log\frac{p_c}{p_{u,v}}+\sum_{(u,v)\in\bar{E}'}\log\frac{1-p_c}{1-p_{u,v}}$$

其中U'、V'、E'为子图H相应的顶点和边，$\log\frac{p_c}{p_{u,v}}$和$\log\frac{1-p_c}{1-p_{u,v}}$分别是相连顶点对和非相连顶点对的权重，$p_{u,v}$定义为在保持每个顶点关联的边数，即顶点的度不变的前提下u和v相连的概率，并取常数$p_c>\max_{(u,v)\in U\times V}p_{u,v}$；(II)寻找$G$中的极大二元组(Biclique)，即$U'$的所有顶点均与$V'$的所有顶点相连的子图；(III)通过逐一增加或去除一个顶点局部优化子图直到子图的得分不再增加，得到最终的稠密子图，即多种癌症表型共享的基因簇(图 1B中椭圆所示)。上述分析在EXPANDER[15]软件中完成。

1.4 癌症共享基因簇的功能分析

采用改进的 Fisher 确切概率检验，即 EASE 打分方法(EASE score)对癌症共享的基因簇进行功能富集分析。为控制假阳性率，应用 Bonferroni 法进行多重检验校正，选取校正后α=0.05 作为显著性水平。本文主要关注基因簇所含基因在基因本体数据库(Gene Otology, GO)中三大分支：生物过程(Biological Process, BP)、分子功能(Molecular Function, MF)和细胞组分(Cellular Component, CC)中的富集情况，探索多效性基因簇在癌症发生发展过程中的作用。基因簇若在 GO 节点得到富集则称为癌症共享的基因功能模块。利用 DAVID 工具(http://david.abcc.ncifcrf.gov/home.jsp)完成富集分析。

1.5 癌症相关网络的构建

基于癌症共享的基因功能模块进一步构建癌症相关网络，了解不同癌症表型之间的遗传关系。若两种癌症共享至少一个基因功能模块，则将它们相连，并计算每个节点(即癌症表型)的接近中心度(Closeness Centrality)和赋予边权重。节点i的接近中心度计算如下[16]：

$$C_i=\sum\nolimits_j\left[d_{ij}\right]^{-1}=\frac{1}{\sum_j d_{ij}}$$

其中，分母表示节点i到所有其他节点$j(j\in V$，V是网络中所有节点的集合)最短路径的和，接近中心度高的节点倾向于位于网络中心。在本文中，接近中心度反映了癌症i与其他癌症表型在分子机制上的相关程度，接近中心度越高则该癌症与越多其他癌症类型存在相关。边的权重w_{ij}是网络中癌症i和癌症j共享基因功能模块的数目，权重越大则两种癌症共享的基因功能模块越多，一定程度上反映了两种癌症之间在分子机制上的相似性。

2 结果与分析

2.1 癌症共享的基因簇

将不同芯片上的探针映射到基因并取交集，共有 12 939 个基因，经过t检验和 logFC 的筛选，最终保留 10 417 个基因表达信息，构建 10417×20 的数据矩阵，其中元素为 logFC 值。应用 SAMBA 算法对该矩阵进行双聚类，得到 22 个癌症共享的基因簇，平均包含 6 种癌症表型和 61 个基因，其中最大的基

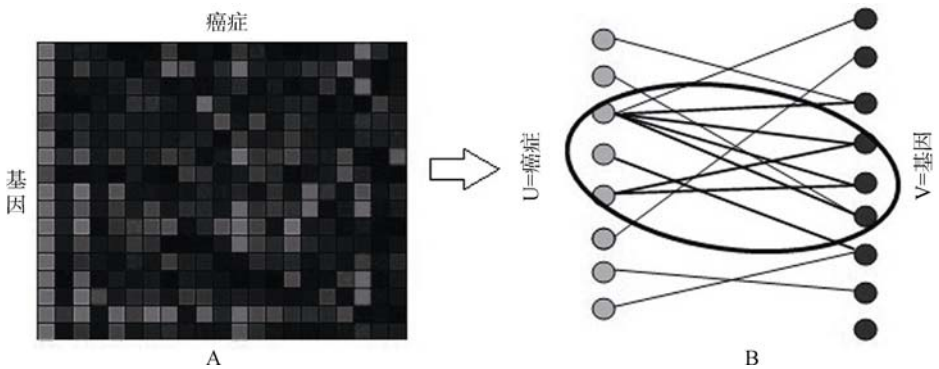

图 1 SAMBA 算法挖掘癌症共享基因簇示意图

A：数据矩阵，元素为基因在癌症类型下的 logFC 值；B：二部图，边表示基因在相应癌症中差异表达。

因簇包括 8 种癌症表型和 101 个基因; 最小的基因簇包括 4 种癌症表型和 14 个基因(表 2)。

2.2　癌症共享的基因功能模块

进一步对 22 个癌症共享的基因簇进行富集分析, 在 Bonferroni 校正后 $P<0.05$ 的检验水准下, 共有 17 个基因簇富集到 GO 数据库中的 BP、MF 和 CC 三大分支的不同节点, 得到富集的基因簇即为癌症共享的基因功能模块(图 2), 与多种生物学机制有关, 主要是细胞有丝分裂(M 期)姐妹染色单体分离的调控、免疫应答和炎症反应、细胞分化、胶原组织、糖代谢等。

如图 2A 所示, 11 个模块富集到深度 $l\geqslant 6$ 的 BP 节点。模块 1、模块 2 和模块 10 主要参与细胞周期尤其是 M 期姐妹染色单体分离的调控, 称为“细胞周期”模块, 显著富集到 spindle checkpoint(模块 1：Bonferroni P=0.001; 模块 2：Bonferroni P= 6.79×10^{-6}; 模块 10：Bonferroni P=0.007)、M phase(模块 1：Bonferroni $P=5.62\times10^{-40}$; 模块 2：Bonferroni $P=2.18\times10^{-36}$; 模块 10：Bonferroni $P=1.07\times10^{-6}$)等生物过程, 3 个模块中的部分基因已知与多种癌症表型发生有关, 如*KIF2C*、*KIF4A*、*KIF20A*、*KIF14*、*KIF11*、*BIRC5*、*BUB1B*、*AURKA*、*CDK1*、*CENP-E*、*CENP-F*等[17~19], 共有 12 种癌症共享这 3 个基因模块(表 2), 均为上皮性肿瘤。模块 5 和模块 19 主要参与免疫应答, 为“免疫应答”模块, 富集到positive regulation of immune response(模块 5：Bonferroni $P=2.54\times10^{-7}$)、immunoglobulin mediated immune response(模块 5：Bonferroni $P=3.89\times10^{-4}$)、complement activation(模块 5：Bonferroni $P=8.31\times10^{-5}$; 模块 19：Bonferroni P=0.010)等生物过程, 部分基因已知可能参与多种癌症的发生发展如*C3*、*HLA-DRB*、*HLA-DMA*[20,21], 本文显示可能是宫颈鳞癌、乳腺癌、胃癌、脂肪肉瘤、肾上腺皮质癌、脑胶质瘤、膀胱癌的共享基因。

表 2　基于双聚类挖掘的癌症共享的基因簇

基因簇	#癌症	#基因	癌症表型
1	5	111	卵巢腺癌、宫颈鳞癌、肺腺癌、大肠癌、膀胱癌
2	8	101	乳腺癌、卵巢腺癌、宫颈鳞癌、胃癌、胰腺癌、大肠癌、间皮瘤、膀胱癌
3	4	94	肺腺癌、鼻咽癌、食管癌、脂肪肉瘤
4	6	91	乳腺癌、宫颈鳞癌、肝细胞性肝癌、皮肤鳞状细胞癌、间皮瘤、肾上腺皮质癌
5	4	88	胃癌、脂肪肉瘤、膀胱癌、肾上腺皮质癌
6	5	83	乳腺癌、宫颈鳞癌、肝细胞性肝癌、黑色素瘤、间皮瘤
7	5	78	宫颈鳞癌、胃癌、肝细胞性肝癌、黑色素瘤、间皮瘤
8	5	73	卵巢腺癌、肺腺癌、胃癌、胰腺癌、黑色素瘤
9	6	60	卵巢腺癌、肺腺癌、胃癌、胰腺癌、肝细胞性肝癌、黑色素瘤
10	9	55	乳腺癌、宫颈鳞癌、肺腺癌、鼻咽癌、胃癌、胰腺癌、食管癌、皮肤鳞状细胞癌、间皮瘤
11	5	45	卵巢腺癌、肺腺癌、胃癌、食管癌、黑色素瘤
12	5	42	卵巢腺癌、肺腺癌、鼻咽癌、胃癌、胰腺癌
13	6	41	卵巢腺癌、肺腺癌、胃癌、肝细胞性肝癌、黑色素瘤、肾透明细胞癌
14	5	41	胃癌、食管癌、黑色素瘤、间皮瘤、脑胶质瘤
15	5	38	宫颈鳞癌、大肠癌、黑色素瘤、间皮瘤、脂肪肉瘤
16	6	30	乳腺癌、卵巢腺癌、胃癌、肝细胞性肝癌、肾透明细胞癌、肾上腺皮质癌
17	6	29	乳腺癌、胃癌、脂肪肉瘤、膀胱癌、肾上腺皮质癌、急性髓细胞性白血病
18	10	25	宫颈鳞癌、乳腺癌、鼻咽癌、胃癌、胰腺癌、肝细胞性肝癌、黑色素瘤、皮肤鳞状细胞癌、间皮瘤、肾透明细胞癌
19	6	24	乳腺癌、宫颈鳞癌、胃癌、脂肪肉瘤、肾上腺皮质癌、脑胶质瘤
20	5	19	胰腺癌、食管癌、肝细胞性肝癌、黑色素瘤、肾透明细胞癌
21	6	17	乳腺癌、肺腺癌、鼻咽癌、食管癌、肾上腺皮质癌、急性髓细胞性白血病
22	4	14	乳腺癌、食管癌、皮肤鳞状细胞癌、脑胶质瘤

模块 7 主要参与炎症反应和维持细胞离子平衡，称为“炎症反应”模块，富集到inflammatory response (Bonferroni P=6.81×10^{-4})、cellular cation homeostasis(Bonferroni P=0.044)，与宫颈鳞癌、胃癌、肝细胞性肝癌、黑色素瘤、间皮瘤 5 种癌症表型有关。模块 8、模块 9 和模块 13 主要参与上皮细胞分化，称为“细胞分化”模块，显著富集到epidermal cell differentiation(模块 8：Bonferroni P=5.72×10^{-9}；模块 9：Bonferroni P=1.84×10^{-9}；模块 13：Bonferroni P=6.61×10^{-11})等生物过程，其中大部分基因的确切功能不清楚，可能通过多种机制与癌症形成或转移有关，如*AHNAK2*、*ANXA1*、*S100A7*、*SPRR2B*、*SPRR1B*等[22~25]，7 种癌症表型共享这些基因(表 2)。此外，模块 3 和模块 10 还参与胶原组织，富集到collagen fibril organization(模块 3：Bonferroni P=0.001；模块 10：Bonferroni P=4.16×10^{-5})，可能与肿瘤的浸润有关[26]。模块 15 称为“糖代谢”模块，富集到 glucose metabolic process(Bonferroni P=6.53×10^{-4})，糖消耗的增加可见于多种肿瘤[27]，模块中部分基因与糖酵解有关，如*PFKFB1*、*GPD1* 等，是 5 种癌症表型的共享基因(表 2)。

如图 2B所示，10 个模块的基因富集到深度 $l\geqslant5$ 的MF节点。模块 1 和模块 2 主要富集到ATP binding(模块 1：Bonferroni P=9.94×10^{-9}；模块 2：Bonferroni P=1.17×10^{-8})、microtubule motor activity (模块 1：Bonferroni P=1.23×10^{-4}；模块 2：P=2.72×10^{-6})。模块 5 和模块 17 富集到MHC class II receptor activity(模块 5：Bonferroni P=7.41×10^{-4}；模块 17：Bonferroni P=2.01×10^{-6})，可能与免疫应答有关。模块 8、模块 9 和模块 11 主要富集到endopeptidase inhibitor activity(模块 8：Bonferroni P=0.048；模块 9：Bonferroni P=0.016)，可能与癌症的侵袭和转移有关，如基因*KLK6*、*PI3*[30~32]。此外，模块 4 富集到dihydrodiol dehydrogenase activity(Bonferroni P=0.044)，富集的基因(*AKR1C1*、*AKR1C2*、*AKR1C3*)属于醛/酮还原酶超家族，其中*AKR1C3* 与细胞生长或分化有关，但与癌症的关系仍不清楚。模块 6 富集到glycosaminoglycan binding(Bonferroni P=0.025)，其中部分基因已知与癌症相关，如*TGFBR3*、*LYVE1*等[28,29]，但该模块与癌症的关系仍需进一步研究。模块 10 富集到platelet-derived growth factor binding(Bonferroni P=9.58×10^{-4})，已知血小板衍生生长因子(platelet-derived growth factor, PDGF)可能与多种癌症表型有关[33]，但该模块和癌症的关系仍需进一步研究。

如图 2C 所示，10 个模块的基因富集到深度 $l\geqslant6$ 的 CC 节点。模块 1、模块 2 和模块 10 富集到 microtubule cytoskeleton(模块 1：Bonferroni P=2.09×10^{-23}；模块 2：Bonferroni P=2.94×10^{-20}，模块 10：Bonferroni P=1.03×10^{-6})、condensed chromosome kinetochore(模块 1：Bonferroni P=1.02×10^{-12}；模块 2：Bonferroni P=6.61×10^{-15}；模块 10：Bonferroni P=0.008)等，即基因主要位于细胞骨架和染色体。模块 5 和模块 17 富集于 MHC class II protein complex(模块 5：Bonferroni P=1.36×10^{-4}；模块 17：Bonferroni P= 1.95×10^{-5}，定位于主要组织相容性复合物。模块 3 富集到 fibrillar collagen(Bonferroni P=1.05×10^{-6})，主要位于细胞外基质。模块 4 和模块 19 富集到 platelet alpha granule(模块 4：Bonferroni P=0.004；模块 19：Bonferroni P=1.33×10^{-4})，主要位于血小板α颗粒。模块 13 和模块 16 富集到 intermediate filament(模块 13：Bonferroni P=0.006；模块 16：Bonferroni P=0.002)，主要位于中间丝蛋白。

2.3 构建癌症相关网络

基于癌症共享的基因功能模块，构建一个有权重的癌症相关网络图，节点大小表示节点接近中心度的高低，线的粗细表示边权重的大小。如图 3 所示，该网络包含了 19 种癌症表型，不同癌症之间共享基因功能模块，反映了癌症在分子机制上紧密的遗传关系，其中，胃癌(接近中心度，C=1.00)、乳腺癌(C=1.00)、宫颈鳞癌(C=0.90)和间皮瘤(C=0.90)处于网络中心，提示与其他癌症在分子机制上相关程度较高；而胃癌与卵巢腺癌(权重，w=6)，宫颈鳞癌和间皮瘤(w=6)，胃癌和黑色素瘤(w=6)，胃癌和乳腺癌(w=5)，胃癌和肺腺癌(w=5)，宫颈鳞癌和乳腺癌(w=5)共享较多的基因功能模块，提示它们两两之间在分子机制上较为相似。结合图 2，我们推测癌症之间的相似性可能与多种因素有关：(I)组织起源，如胃癌和乳腺癌都是腺癌；(II)致癌机制，如脏膜长期受石棉刮擦刺激，宫颈长期受性病的刺激，造成损伤、修复的反复，促进癌变为间皮瘤或宫颈癌；又如

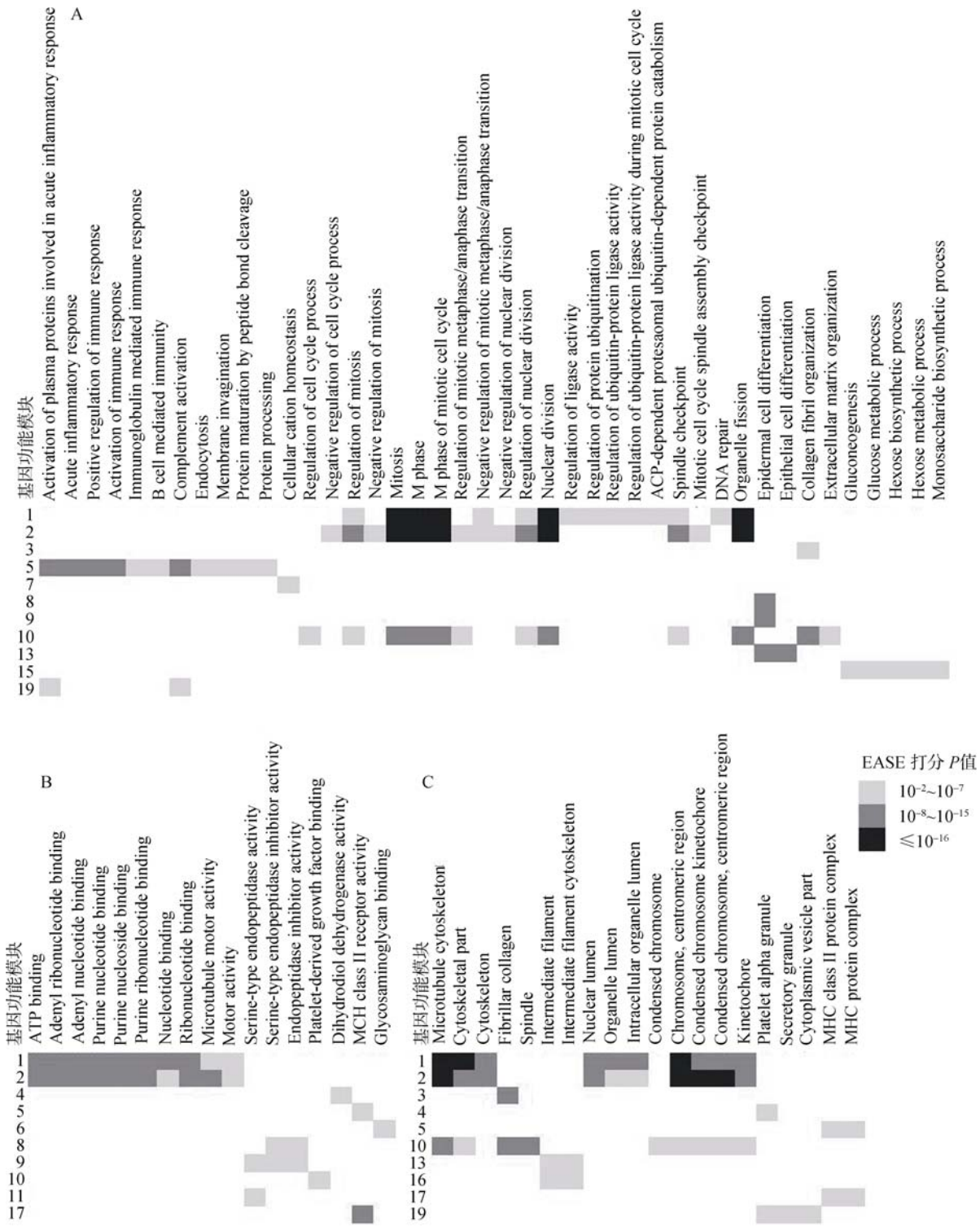

图 2　基因功能模块富集到的 **GO** 节点

A：富集的 BP(Biological Process)节点; B：富集的 MF(Molecular Function)节点; C：富集的 CC(Cellular Component)节点。行：基因功能模块; 列：富集的 BP、MF、CC 节点; 颜色深浅表示富集程度高低, 即 EASE 打分方法未校正的 *P* 值。

宫颈癌和乳腺癌可能与性激素有关等。此外, 癌症之间的相似性还可能与其侵袭和转移特性有关。特别的, 急性髓细胞性白血病只聚集于一个基因功能模块, 而多发性骨髓瘤均不在任一基因功能模块聚

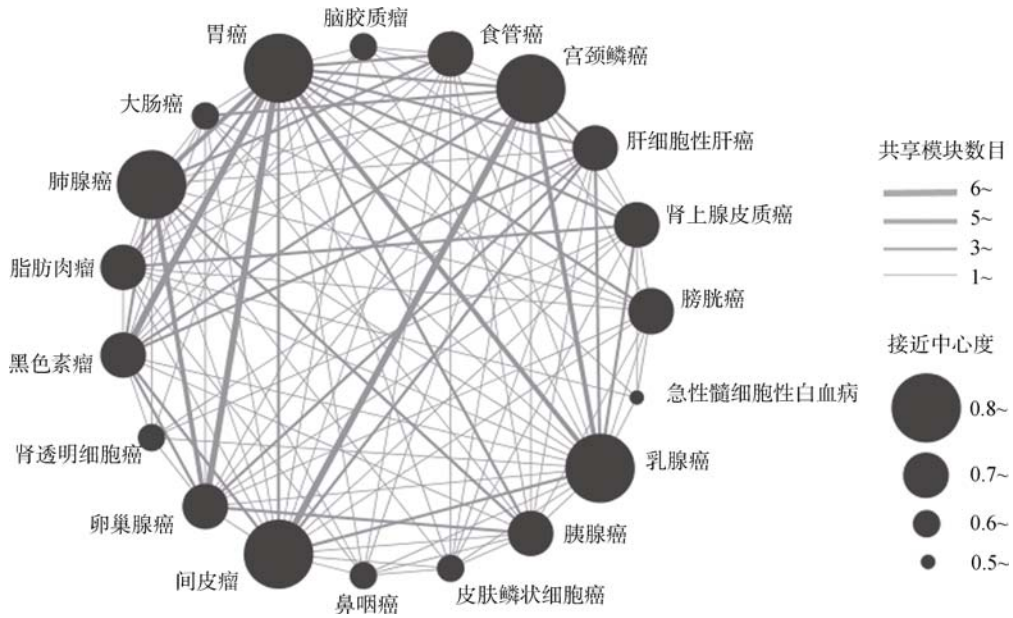

图 3 基于基因功能模块构建的癌症相关网络图

集，提示分子机制和其他癌症类型的差异较大。

3 讨 论

SAMBA算法作为一种双聚类方法，有效利用矩阵的局部信息，适用于高维数据信息的挖掘，已经被用于构建酵母的基因调控网络[13]。在具有相关功能的基因较大可能影响相同或相似表型的假设下，本文采用SAMBA算法挖掘与不同癌症表型发生发展相关的共享分子机制。不同于以往的的研究，本文综合多种常见癌症表型的基因表达数据，从整体上分析癌症共同的分子机制，有利于建立微效基因和癌症之间的联系；其次本文基于基因功能模块而非基因，建立起癌症表型之间的联系，结合富集分析有利于探索癌症共享基因的相关的生物学机制。基因功能模块提供了癌症可能的普遍性分子机制，支持了基因多效性在癌症中的普遍性的观点，即癌症的基因变异并不是特有的，许多基因变异普遍存在于许多癌症中[34]，也支持了基因调控网络是模块化组织方式，模块的功能的失调可能导致癌症的观点。

有丝分裂姐妹染色单体分离的调控是模块 1、模块 2 和模块 9 主要的生物学机制。通常细胞周期沿着G_1-S-G_2-M期有序运转，完成DNA合成和有丝分裂，为了确保这一过程的准确性，细胞在分裂过程中存在广泛的防御机制，如纺锤体检查点，M期动粒与纺锤体不恰当的连接可触发该检查点，延缓有丝分裂进入后期。模块中部分基因已知参与纺锤体检查点活性的调节，如*CDC20* 阻断后期促进因子(Anaphase promoting complex, APC)的激活，*CENP-E* 参与激活*BUBR1*，均能上调检查点活性[35]；*CENP-F* 则参与检查点的去激活；磷酸化*CDK1* 也能激活APC而使有丝分裂进入后期等[36]。防御机制缺陷将使DNA损伤积聚，导致基因组不稳定。此外，驱动蛋白超家族(Kinesin superfamily proteins, KIFs)参与纺锤体的形成、染色体和细胞核的移动等，KIFs失调也将导致遗传物质的不均匀分离，破坏基因组稳定性[17]。基因组的不稳定性可能是肿瘤逐渐获得抵制凋亡，永生化和逃避细胞周期调控这些疾病特征的一个重要途径，是多种癌症的共同特征[37]。模块 4、模块 8 和模块 11 的主要生物学机制是上皮细胞分化，其中基因致癌机制并不清楚，但以往研究提示模块中的部分基因有微弱或潜在的致癌效应，如 *AHNAK2* 维持细胞钙离子平衡，在多种癌细胞中低表达；*ANXA1* 有抗炎效应，不同癌症细胞具有不同的表达模式[38]；*SPRR2B*与癌症淋巴结转移有关[24]，*SPRR1B*过表达影响细胞的正常分裂[25]等。癌症共享分子机制还与体液免疫和炎症反应有关(模块 13、15 和 18)，已有研究提示体液免疫在肿瘤的发生发展中发挥双重作用，一方面进行免疫监视，清除癌变细胞，一方面可能通过上调细胞增殖信号和抑制癌变细胞凋亡促进肿瘤的生长[39,40]。此外，癌症共

享分子机制还与癌症具有浸润转移的特性(模块 9 和模块 10)和增加糖消耗维持癌细胞的生长优势有关(模块 20)。

基于基因功能模块，本文建立了癌症之间的遗传联系，一方面提示尽管癌症的组织来源和行为特点均不同，但在分子机制上存在复杂紧密的联系；另一方面提示癌症之间的联系可能不仅与组织起源有关[16]，还可能与共同致癌机制(如长期的刺激和激素)和癌症生物特性(如转移)有关。癌症相关网络图提示两类血液系统疾病(急性髓细胞白血病和多发性骨髓瘤)的分子机制可能与其它癌症类型存在较大差异；也可能受本文未能识别*TP53*、*PTEN*等常见抑癌基因的影响所致。

本文基因功能模块中癌症与基因的关系部分得到了以往研究的支持，但基因的表达和癌症是否存在因果关系仍需要进一步的研究加以验证。另外，即使是同一种癌症，不同的数据集包含不同的癌症亚型或分期，可能对本研究结果产生一定的影响。随着公共生物数据的积累，进一步获得癌症的不同亚型或分期的表达数据成为可能，若对癌症病例进一步分类，得到不同亚型/分期下癌症的表达值，或增加分析的癌症数据集，可能会使结果更全面和准确。总而言之，本文应用双聚类算法挖掘癌症共享的基因功能模块，为进一步揭示癌症共同致病机制提供了线索，结合基因簇的功能分析，有助于解释多效性基因在致癌过程中所起的作用；此外，基因功能模块预测了可能的癌症和基因间的关系，有助于进一步剖析癌症复杂的遗传机制，对复杂性疾病的预防、诊断及治疗具有重要意义。

参考文献**(References)**:

[1] Yu W, Clyne M, Khoury MJ, Gwinn M. Phenopedia and Genopedia: disease-centered and gene-centered views of the evolving knowledge of human genetic associations. *Bioinformatics*, 2010, 26(1): 145–146. DOI

[2] Stearns FW. One hundred years of pleiotropy: a retrospective. *Genetics*, 2010, 186(3): 767–773. DOI

[3] Becker KG. The common variants/multiple disease hypothesis of common complex genetic disorders. *Med Hypotheses*, 2004, 62(2): 309–317. DOI

[4] Naccarati A, Polakova V, Pardini B, Vodickova L, Hemminki K, Kumar R, Vodicka P. Mutations and polymorphisms in TP53 gene-an overview on the role in colorectal cancer. *Mutagenesis*, 2012, 27(2): 211–218. DOI

[5] Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. *Proc Natl Acad Sci USA*, 2007, 104(21): 8685–8690. DOI

[6] Sivakumaran S, Agakov F, Theodoratou E, Prendergast JG, Zgaga L, Manolio T, Rudan I, McKeigue P, Wilson JF, Campbell H. Abundant pleiotropy in human complex diseases and traits. *Am J Hum Genet*, 2011, 89(5): 607–618. DOI

[7] 姚晨，张敏，邹金凤，李红东，王栋，朱晶，郭政. 可识别多种癌症的基因功能模块. 中国科学 C辑：生命科学, 2009, 39(11): 1092–1096. DOI

[8] Gu X. Evolutionary framework for protein sequence evolution and gene pleiotropy. *Genetics*, 2007, 175(4): 1813–1822. DOI

[9] 朱晶，沈晓沛，肖会，张杨，王靖，郭政. 基于共进化基因功能模块发现候选癌基因. 遗传, 2010, 32(7): 694–700. DOI

[10] Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. *Nucl Acids Res*, 2003, 31(4): e15. DOI

[11] Gautier L, Cope L, Bolstad BM, Irizarry RA. affy-analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 2004, 20(3): 307–315. DOI

[12] Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 2004, 3(1): Article3. DOI

[13] Tanay A, Sharan R, Kupiec M, Shamir R. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc Natl Acad Sci USA*, 2004, 101(9): 2981–2986. DOI

[14] Tanay A, Sharan R, Shamir R. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 2002, 18(Suppl. 1): S136–S144. DOI

[15] Shamir R, Maron-Katz A, Tanay A, Linhart C, Steinfeld I, Sharan R, Shiloh Y, Elkon R. *EXPANDER*-an integrative program suite for microarray data analysis. *BMC Bioinformatics*, 2005, 6: 232, doi: 10.1186/1471-2105-6-232. DOI

[16] Li WT, Wang R, Bai LF, Yan ZM, Sun ZR. Cancer core modules identification through genomic and transcriptomic changes correlation detection at network level. *BMC Syst Biol*, 2012, 6: 64, doi: 10.1186/1752-0509-6-64. DOI

[17] Yu Y, Feng YM. The role of kinesin family proteins in tumorigenesis and progression: potential biomarkers and molecular targets for cancer therapy. *Cancer*, 2010, 116(22): 5150–5160. DOI

[18] Liu ZJ, Ling K, Wu X, Cao J, Liu B, Li SY, Si Q, Cai Y, Yan C, Zhang Y, Weng YG. Reduced expression of cenp-e in human hepatocellular carcinoma. *J Exp Clin Cancer Res*, 2009, 28: 156, doi:10.1186/1756-9966-28-156. DOI

[19] Brown HK, Ottewell PD, Coleman RE, Holen I. The kinetochore protein Cenp-F is a potential novel target for zoledronic acid in breast cancer cells. *J Cell Mol Med*,

2011, 15(3): 501–513. DOI

[20] Dowling P, Clarke C, Hennessy K, Torralbo-Lopez B, Ballot J, Crown J, Kiernan I, O'Byrne KJ, Kennedy MJ, Lynch V, Clynes M. Analysis of acute-phase proteins, AHSG, C3, CLI, HP and SAA, reveals distinctive expression patterns associated with breast, colorectal and lung cancer. *Int J Cancer*, 2012, 131(4): 911–923. DOI

[21] Wolkersdorfer T, Füssel M, Kiesslich T, Neureiter D, Berr F, Aust D, Wolkersdorfer GW. MHC class II genotype- and MHC class I and II phenotype-related parameters in sporadic colorectal cancer. *Oncol Rep*, 2011, 26(5): 1165–1171. DOI

[22] Kouno M, Kondoh G, Horie K, Komazawa N, Ishii N, Takahashi Y, Takeda J, Hashimoto T. Ahnak/Desmoyokin is dispensable for proliferation, differentiation, and maintenance of integrity in mouse epidermis. *J Invest Dermatol*, 2004, 123(4): 700–707. DOI

[23] Nasser MW, Qamri Z, Deol YS, Ravi J, Powell CA, Trikha P, Schwendener RA, Bai XF, Shilo K, Zou X, Leone G, Wolf R, Yuspa SH, Ganju RK. S100A7 enhances mammary tumorigenesis through upregulation of inflammatory pathways. *Cancer Res*, 2012, 72(3): 604–615. DOI

[24] Pasini FS, Maistro S, Snitcovsky I, Barbeta LP, Rotea Mangone FR, Lehn CN, Walder F, Carvalho MB, Brentani MM, Federico MH. Four-gene expression model predictive of lymph node metastases in oral squamous cell carcinoma. *Acta Oncol*, 2012, 51(1): 77–85. DOI

[25] Tesfaigzi Y, Wright PS, Belinsky SA. SPRR1B overexpression enhances entry of cells into the G0 phase of the cell cycle. *Am J Physiol Lung Cell Mol Physiol*, 2003, 285(4): L889–L898. DOI

[26] Ruangpanit N, Chan D, Holmbeck K, Birkedal-Hansen H, Polarek J, Yang CL, Bateman JF, Thompson E W. Gelatinase A (MMP-2) activation by skin fibroblasts: dependence on MT1-MMP expression and fibrillar collagen form. *Matrix Biol*, 2001, 20(3): 193–203. DOI

[27] Gatenby RA, Gillies RJ. Why do cancers have high aerobic glycolysis? *Nat Rev Cancer*, 2004, 4(11): 891–899. DOI

[28] Lin X, Chen YG, Meng AM, Feng XH. Termination of TGF-β superfamily signaling through SMAD dephosphorylation-a functional genomic view. *J Genet Genomics*, 2007, 34(1): 1–9. DOI

[29] Du Y, Liu YW, Wang YZ, He YQ, Yang CX, Gao F. LYVE-1 enhances the adhesion of HS-578T cells to COS-7 cells via hyaluronan. *Clin Invest Med*, 2011, 34(1): E45–E54. DOI

[30] Seiz L, Dorn J, Kotzsch M, Walch A, Grebenchtchikov NI, Gkazepis A, Schmalfeldt B, Kiechle M, Bayani J, Diamandis EP, Langer R, Sweep FC, Schmitt M, Magdolen V. Stromal cell-associated expression of kallikrein-related peptidase 6 (KLK6) indicates poor prognosis of ovarian cancer patients. *Biol Chem*, 2012, 393(5): 391–401. DOI

[31] Hoskins E, Rodriguez-Canales J, Hewitt SM, Elmasri W, Han J, Han S, Davidson B, Kohn EC. Paracrine SLPI secretion upregulates MMP-9 transcription and secretion in ovarian cancer cells. *Gynecol Oncol*, 2011, 122(3): 656–662. DOI

[32] Clauss A, Ng V, Liu J, Piao HY, Russo M, Vena N, Sheng Q, Hirsch MS, Bonome T, Matulonis U, Ligon AH, Birrer MJ, Drapkin R. Overexpression of elafin in ovarian carcinoma is driven by genomic gains and activation of the nuclear factor kappaB pathway and is associated with poor overall survival. *Neoplasia*, 2010, 12(2): 161–172. DOI

[33] 李征, 何剪太. 血清血小板衍生生长因子测定在癌症诊断中的应用价值. 中国现代医学杂志, 2011, 21(17): 2072–2076. DOI

[34] Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, Barretina J, Boehm JS, Dobson J, Urashima M, Mc Henry KT, Pinchback RM, Ligon AH, Cho YJ, Haery L, Greulich H, Reich M, Winckler W, Lawrence MS, Weir BA, Tanaka KE, Chiang DY, Bass AJ, Loo A, Hoffman C, Prensner J, Liefeld T, Gao Q, Yecies D, Signoretti S, Maher E, Kaye FJ, Sasaki H, Tepper JE, Fletcher JA, Tabernero J, Baselga J, Tsao MS, Demichelis F, Rubin MA, Janne PA, Daly MJ, Nucera C, Levine RL, Ebert BL, Gabriel S, Rustgi AK, Antonescu CR, Ladanyi M, Letai A, Garraway LA, Loda M, Beer DG, True LD, Okamoto A, Pomeroy SL, Singer S, Golub TR, Lander ES, Getz G, Sellers WR, Meyerson M. The landscape of somatic copy-number alteration across human cancers. *Nature*, 2010, 463(7283): 899–905. DOI

[35] Li M, Fang X, Wei ZB, York JP, Zhang PM. Loss of spindle assembly checkpoint-mediated inhibition of Cdc20 promotes tumorigenesis in mice. *J Cell Biol*, 2009, 185(6): 983–994. DOI

[36] Chow JPH, Poon RY, Ma HT. Inhibitory phosphorylation of cyclin-dependent kinase 1 as a compensatory mechanism for mitosis exit. *Mol Cell Biol*, 2011, 31(7): 1478–1491. DOI

[37] Kops GJL, Weaver BAA, Cleveland DW. On the road to cancer: aneuploidy and the mitotic checkpoint. *Nat Rev Cancer*, 2005, 5(10): 773–785. DOI

[38] Kang H, Ko J, Jang SW. The role of annexin A1 in expression of matrix metalloproteinase-9 and invasion of breast cancer cells. *Biochem Biophys Res Commun*, 2012, 423(1): 188–194. DOI

[39] Shishido SN, Varahan S, Yuan K, Li XD, Fleming SD. Humoral innate immune response and disease. *Clin Immunol*, 2012, 144(2): 142–158. DOI

[40] Ye XZ, Yu SC, Bian XW. Contribution of myeloid-derived suppressor cells to tumor-induced immune suppression, angiogenesis, invasion and metastasis. *J Genet Genomics*, 2010, 37(7): 423–430. DOI