生 物 工 程 学 报
Chinese Journal of Biotechnology
http://journals.im.ac.cn/cjbcn
DOI: 10.13345/j.cjb.210137

Nov. 25, 2021, 37(11): 4111-4123

·生物技术与方法·

基于集成的共表达网络分析方法研究 3 种癌症的肿瘤相关模块

王梦男[1]，韩明飞[1]，刘炳辉[2]，田春艳[1]，朱云平[1]

1 军事科学院军事医学研究院生命组学研究所 国家蛋白质科学中心 (北京) 蛋白质组研究中心 蛋白质组学国家重点实验室，北京 102206

2 军事兽医研究所，吉林 长春 130000

王梦男，韩明飞，刘炳辉，等. 基于集成的共表达网络分析方法研究 3 种癌症的肿瘤相关模块. 生物工程学报，2021, 37(11): 4111-4123.
Wang MN, Han MF, Liu BH, et al. Integration-based co-expression network analysis to investigate tumor-associated modules across three cancer types. Chin J Biotech, 2021, 37(11): 4111-4123.

摘 要：在肿瘤/癌旁基因表达数据中，差异表达 (DE，differential expression) 代表各种生物条件下基因表达水平的变化，而差异共表达 (DC，differential co-expression) 代表基因对之间相关系数的变化。单独的 DC 和 DE 研究方法已经被广泛应用于人类疾病研究中。但是，目前仍然缺乏有效整合 DC 和 DE 的分析方法。文中提出一个新颖的分析框架 DC&DEmodule，该框架可以基于共表达模块整合 DC 和 DE 的特征，并同时整合多个肿瘤/癌旁表达谱的信息，用以识别与疾病相关的基因共表达模块，包括激活模块 (肿瘤样本中上调且共表达增强) 和失能模块 (肿瘤样本中下调且失去共表达)。将该框架用于分析肝癌、胃癌和结直肠癌各两组微阵列数据，分别得到肝癌、胃癌和结直肠癌的 2、5 和 2 个激活模块以及 5、5 和 1 个失能模块。富集分析表明与同类方法相比，文中的方法在检测已知的肿瘤相关通路和发现新通路方面均具有更高的灵敏度。然后，进一步从这 3 种癌症的激活模块中鉴定出 17、69 和 11 个模块关键基因，其中包含 53 个已报道的预后生物标志物以及 3 个分别与 3 种癌症存活率显著相关的新预后标志物。基于关键基因训练了 3 种癌症的随机森林模型，用于区分 TCGA(The Cancer Genome Atlas) 和 GEO (Gene Expression Omnibus) 数据库中的肿瘤和癌旁样本，结果显示其分类的平均准确性达到了 93%。三种癌症的比较为不同癌症的共有和组织特异性机制提供了新的见解。一系列评估表明，DC&DEmodule 框架能够整合公共数据库中快速积累的表达谱，发现更多疾病中功能失调的生物过程。

关键词：癌症，差异共表达，差异表达，共表达模块，基于集成的分析框架

Received: February 10, 2021; **Accepted:** March 11, 2021
Supported by: Open Project Program of the State Key Laboratory of Proteomics (No. SKLP-O2020005), National Key Research and Development Program of China (No. 2016YFB0201702), National High Technology Research and Development Program of China (863 Program) (No. 2012AA020409), National Basic Research Program of China (973 Program) (No. 2011CB910601).
Corresponding author: Yunping Zhu. Tel: +86-10-61777058; E-mail: zhuyunping@ncpsb.org.cn
蛋白质组学国家重点实验室开放课题 (No. SKLP-O2020005)，国家重点研发计划 (No. 2016YFB0201702)，国家高技术研究发展计划 (863 计划) (No. 2012AA020409)，国家重点基础研究发展计划 (973 计划) (No. 2011CB910601) 资助。

Integration-based co-expression network analysis to investigate tumor-associated modules across three cancer types

Mengnan Wang[1], Mingfei Han[1], Binghui Liu[2], Chunyan Tian[1], and Yunping Zhu[1]

1 *State Key Laboratory of Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences* (*Beijing*), *Institute of Lifeomics, Academy of Military Medical Sciences, Academy of Military Sciences, Beijing* 102206, *China*

2 *Institute of Military Veterinary Medicine, Academy of Military Medical Sciences, Academy of Military Sciences, Changchun* 130000, *Jilin, China*

Abstract: In case/control gene expression data, differential expression (DE) represents changes in gene expression levels across various biological conditions, whereas differential co-expression (DC) represents an alteration of correlation coefficients between gene pairs. Both DC and DE genes have been studied extensively in human diseases. However, effective approaches for integrating DC–DE analyses are lacking. Here, we report a novel analytical framework named DC&DEmodule for integrating DC and DE analyses and combining information from multiple case/control expression datasets to identify disease-related gene co-expression modules. This includes activated modules (gaining co-expression and up-regulated in disease) and dysfunctional modules (losing co-expression and down-regulated in disease). By applying this framework to microarray data associated with liver, gastric and colon cancer, we identified two, five and two activated modules and five, five and one dysfunctional module(s), respectively. Compared with the other methods, pathway enrichment analysis demonstrated the superior sensitivity of our method in detecting both known cancer-related pathways and those not previously reported. Moreover, we identified 17, 69, and 11 module hub genes that were activated in three cancers, which included 53 known and three novel cancer prognostic markers. Random forest classifiers trained by the hub genes showed an average of 93% accuracy in differentiating tumor and adjacent normal samples in the TCGA and GEO database. Comparison of the three cancers provided new insights into common and tissue-specific cancer mechanisms. A series of evaluations demonstrated the framework is capable of integrating the rapidly accumulated expression data and facilitating the discovery of dysregulated processes.

Keywords: cancer, differential co-expression, differential expression, gene co-expression module, integration-based analytical framework

基因往往不是独立发挥作用，而是多个基因相互协调，共同参与某一生物学过程。参与同一生物学过程的基因对或基因模块在不同样本中同时表现出上调或下调的趋势，称之为基因共表达现象。在此基础上，研究人员将疾病与健康状态相比新建立或失去共表达现象的基因模块，称为差异共表达 (Differential co-expression，DC) 基因模块。在疾病中，新建立的共表达模块往往参与因疾病而激活的生物过程，而失去共表达趋势的模块往往参与在疾病中失调的生物过程。近年来，差异共表达分析已被广泛用于探索各类疾病的机制[1-10]。此外，差异表达 (Differential expression，DE) 可以鉴定疾病与健康状态下表达水平显著变化的基因，因此一直是在分子水平上研究人类致病基因的重要手段。DC 基因和 DE 基因均被证实与疾病的发生发展相关，但目前大部分研究仍集中在分别鉴定 DC 和 DE 基因[11-17]，DC 和 DE 分析未能得到有效的整合。尽管 Lui 等提出的 DECODE 方法可以整合单个基因的 DC 和 DE 特征来鉴定致病基因[18]，但目前仍缺乏有效整合共表达模块的 DC 和 DE 基因特征的方法。

针对这一问题，我们开发了一种新的整合 DC 和 DE 特征的分析框架 DC&DEmodule (图 1)，并用于分析 Gene Expression Omnibus (GEO)数据库中下载的 6 组肿瘤/癌旁微阵列 (Microarray) 数据(肝癌、胃癌和结直肠癌各两组)。DC&DEmodule

包含两个关键步骤：1) 分别鉴定肿瘤和癌旁组织的共表达基因模块；2) 整合 DC 和 DE 特征鉴定肿瘤相关模块，包括激活模块 (肿瘤样本中上调且共表达增强) 和失能模块 (肿瘤样本中下调且失去共表达)。

最终，我们在肝癌、胃癌和结肠癌中分别确定了 2、5 和 2 个激活模块以及 5、5 和 1 个失能模块，并且在 3 种癌症的激活模块中分别鉴定出 17、69 和 11 个关键基因。这些关键基因包含 53 个已报道的和 3 个新的癌症预后标志物 TCEB1

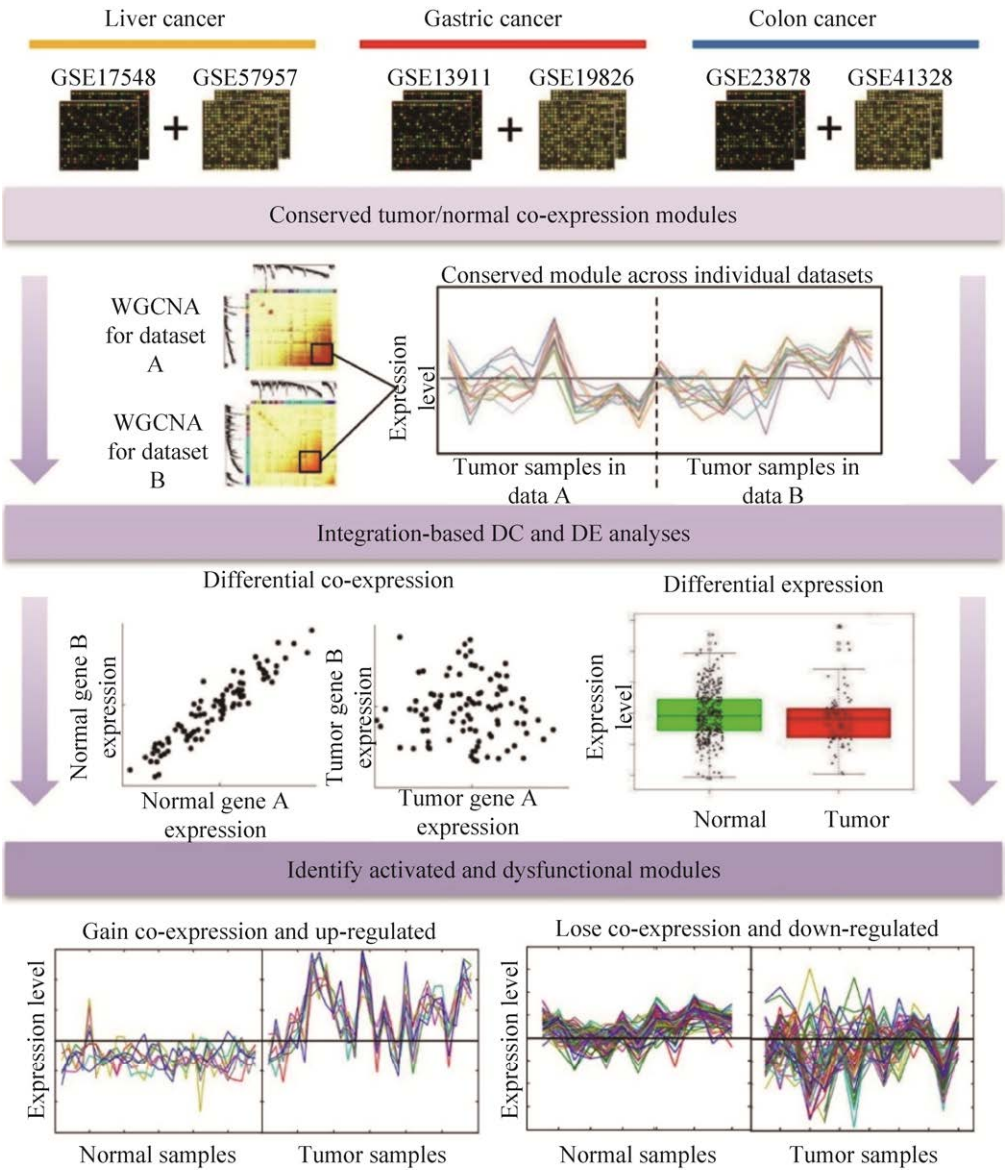

图 1　肿瘤相关模块鉴定的工作流程

Fig. 1　Schematic procedure for the identification of tumor-associated modules. We integrated two GEO microarray datasets associated with liver, gastric and colon cancer, respectively. For each cancer type, we used WGCNA to identify tumor/normal co-expression modules in each of the two datasets, then extracted the significantly conserved modules across individual datasets. We then identified the modules showing both differential expression levels and distinct co-expression strengths between normal and tumor conditions, including oncogenic modules (gaining co-expression and up-regulated in tumor) and dysfunctional modules (losing co-expression and down-regulated in tumor).

(transcription elongation factor B (SIII), polypeptide 1 (15 kDa, elongin C)、RFC4 (Replication Factor C Subunit 4)和 TRPC4AP(Transient Receptor Potential Cation Channel Subfamily C Member 4)。生存分析表明 3 个新标志物分别与肝癌、胃癌和结直肠癌的总体生存时间 (Overall survival) 显著相关。随后，我们分别基于 3 种癌症的关键基因训练了 3 个随机森林分类器，用于区分 TCGA 和 GEO 数据库中的肿瘤和癌旁样本，结果显示分类的平均准确性达到 93%。最后，我们利用独立的 DC 分析 (基因集共表达分析，GSCA(Gene Set Co-expression analysis)、独立的 DE 分析 (Student's *t*-test)以及现有整合 DC 和 DE 分析的工具 (DECODE) 分析 3 种癌症的表达谱，并与 DC&DEmodule 的结果比较。通路富集分析表明 DC&DEmodule 在检测已报道的肿瘤相关通路和发现新通路方面均具有更高的灵敏度。综上，我们的整合分析框架可以更好地发现肿瘤相关通路和生物标志物，并为 DC 和 DE 的联合分析提供新的视角。

1 材料与方法

1.1 材料

从 GEO 数据库中(www.ncbi.nlm.nih.gov/geo)分别收集肝癌、胃癌和结直肠癌各两个微阵列表达谱。所有表达谱均含有从同一患者的肿瘤和相邻的癌旁组织切除的样本，包括 GSE57957[19]、GSE17548[20] 、GSE13911[21] 、GSE19826[22] 、GSE23878[23]和 GSE41328[24]。根据以下流程对原始数据进行预处理：1) 进行 RMA[25]数据标准化；2) 删除在所有样本中表达量空值比例大于 20%的基因 (当多个探针匹配到同一基因时，优先保留最大的表达值)；3) 选取在同一癌症的两组微阵列数据中共同检测到的基因；4) 过滤掉在 neXtprot[26]中不存在的基因。数据预处理结果见补充表 S1。

1.2 方法

1.2.1 构建基因共表达网络

应用 R 语言中 WGCNA 程序包对每个表达谱分别构建肿瘤和癌旁样本共表达网络 (图 2 步骤 1)，接着对网络进行层次聚类，然后对层次聚类树进行动态切割[27]，将具有相似表达模式的基因归为一个共表达模块。为了整合每种癌症类型的两个独立的表达数据，我们取两组数据的所有模块组成两两模块对，使用费舍尔精确检验评估每个模块对中基因的保守性。当一个模块对的费舍尔精确检验 P 值小于 0.05，且具有多于 10 个重叠基因时，这些重叠基因就组成两组数据的一个保守模块。最终我们鉴定到每种癌症高度保守的肿瘤/癌旁模块。模块结果见补充表 S2。

1.2.2 差异共表达分析

首先利用 R 语言中 GSCA 程序包[15]识别肿瘤和癌旁样品之间显著的 DC 模块。对于一个包含 n 个基因的共表达模块 M，计算其两两基因组成的 $P_M(P_M=C_n^2)$ 个基因对的相关系数 $\rho_p(p=1,...,P_M)$，然后利用欧几里得距离计算模块 M 在一个表达谱 S_1 中肿瘤和癌旁的 DC 分数 $dc_M=\sqrt{\frac{1}{P_M}\sum_{p=1}^{P_M}(\rho_p^T-\rho_p^{NT})^2}$, $p=1,...,P_M$。继而计算两个独立表达谱 S_1 和 S_2 的合并 DC 分数 $DC_M=\sqrt{\frac{1}{P_M}[\sum_{p=1}^{P_M}(\rho_p^{T_{S_1}}-\rho_p^{NT_{S_1}})^2+\sum_{p=1}^{P_M}(\rho_p^{T_{S_2}}-\rho_p^{NT_{S_2}})^2]}$, p=1,…,P_M。

应用置换检验识别显著 DC 模块，置换检验的 P 值为 $[1+\sum_{b-1}^{B-1}I(DC_M^b>DC_M)]/B$，($B$=9 999)，$DC_M^b$ 为 9 999 次随机置换肿瘤和癌旁样本标签得到的 DC 打分。然后利用 Benjamini-Hochberg 方法对 P 值进行多重假设检验校正[28]。最终，我们分别在肿瘤和癌旁组织的共表达模块中鉴定到两类显著 DC 模块。其中，显著 DC 的肿瘤模块表示基因在肿瘤组织新建共表达，显著 DC 的癌旁模块表示基因在癌旁中共表达而在肿瘤组织中丧失共表达 (图 2 步骤 2)。

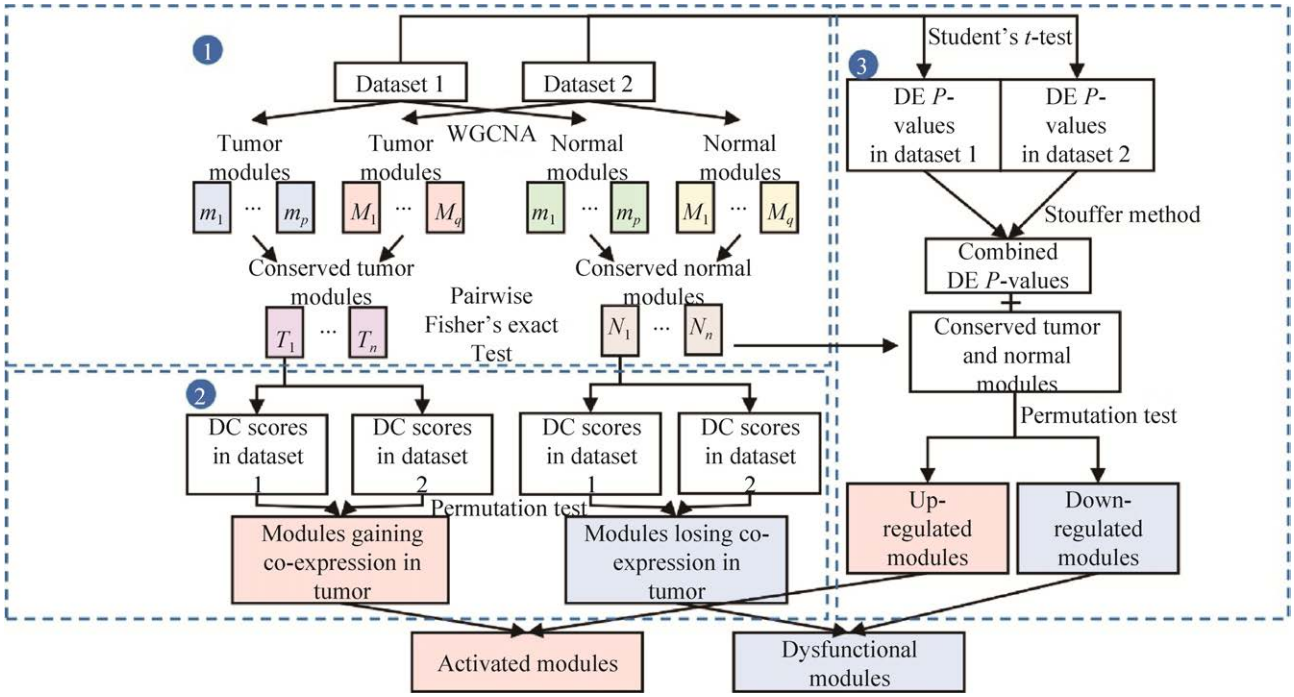

图 2　DC-DE 整合分析框架的工作流程

Fig. 2　Flowchart of the DC–DE integration analytical framework. Using the analysis of two cancer datasets as an example, we first identified tumor and normal co-expression modules for each dataset, then extracted the significantly conserved tumor/normal modules across individual datasets using pairwise Fisher's exact test (step 1). For each significantly conserved module, we performed integration-based DC (step 2) and DE (step 3) analyses to identify modules gaining co-expression, losing co-expression, up-regulated, and down-regulated in the tumor, respectively. Modules gaining co-expression and up-regulated in tumor were identified as activated modules and those losing co-expression and down-regulated in tumor were identified as dysfunctional modules.

接着我们鉴定显著 DC 模块内的关键基因，对于一个包含 n 个基因的显著 DC 模块 M，计算其中每个基因 g 的差异共表达分数 $DC_g=\sqrt{\frac{1}{n-1}[\sum_{p=1}^{n-1}(\rho_p^{T_{S_1}}-\rho_p^{N_{S_1}})^2+\sum_{p=1}^{n-1}(\rho_p^{T_{S_2}}-\rho_p^{N_{S_2}})^2]}$，其中 ρ_p 代表基因 g 和模块 M 中其他 $(n–1)$个基因的皮尔森相关系数，T_{s1}、N_{s1}、T_{s2} 和 N_{s2} 分别代表两个独立的表达谱 S_1 和 S_2 的肿瘤和癌旁样本。最后应用置换检验鉴定显著 DC 模块中的关键基因(置换检验 $P<0.05$)。

1.2.3　差异表达分析

首先基于 Student's *t*-test 计算每组独立的微阵列表达谱中基因的差异表达 P 值，然后对 T 检验 P 值进行转换：当基因在肿瘤中上调时 P 值转化为 $p'=P/2$，下调时转换为 $p'=1–P/2$。因此当 p' 越接近 0 时，基因在肿瘤中越显著上调，越接近 1 时，基因在肿瘤中越显著下调。我们使用斯托弗法 (Stouffer method) 合并两个独立微阵列表达谱中同一基因的转换 P 值为合并 P 值 $S=1/\sqrt{2}\times[\psi^{-1}(P_1')+\psi^{-1}(P_2')]$ [29-30]，其中 P_1' 和 P_2' 表示同一基因在两组独立表达谱的转换 P 值，ψ 表示标准正态分布函数。通过随机置换样本标签产生 B (B=10 000) 个随机的合并 DE 打分 $S^{1\sim B}=1/\sqrt{2}\times[\psi^{-1}(RP_1')+\psi^{-1}(RP_2')]$，然后利用置换检验评估合并的 DE 分数 S 的显著性 ($P_{combined}=[1+\sum_{b=1}^{B-1}I(S^b>S)/B]$) [31]。然后利用 Benjamini-Hochberg 方法对 P 值进行多重假设检验校正[28]。根据 P'的特征可以推导出合并 P 值越接近 0 时，基因越显著上调，合并 P 值越接近 1

时，基因越显著下调 (图 2 步骤 3)。基于合并 *P* 值可以进一步计算忽略上调/下调的合并 DE 分数 $DES=|P-0.5|$，*DES* 越大基因差异表达越显著。

紧接着我们识别上下调的基因模块，对于一个包含 *n* 个基因的共表达模块 *M*，首先计算其 DE 打分即 $DE_M=\sqrt{\frac{1}{n}\sum_{i=1}^{n}DES_i^2}$，其中 DES_i 代表基因的合并 DE 打分。随后在全部基因中随机选择的 *n* 个基因，重复 *B* 次 (*B*=10 000) 得到 *B* 个随机模块，并计算它们的随机 DE 打分 DE_M^{1-B}。如果 DE_M 高于 95% DE_M^{1-B} ($P<0.05$)，模块 *M* 即为显著差异表达的模块。然后我们确认模块 *M* 的差异表达方向。方法是计算模块 *M* 中 *n* 个基因的合并差异表达 *P* 值 ($P_{combined}$) 的均值，即 $Avg_M=\sqrt{\frac{1}{n}\sum_{i=1}^{n}P_i^2}$。根据上述可知 Avg_M 越大模块上调越显著，越小模块下调越显著。然后利用置换检验评估 Avg_M 的显著性，当 Avg_M 大于 95%随机模块的结果 Avg_M^{1-B} 时模块 *M* 显著上调，当 Avg_M 小于 95%随机模块的结果时模块 *M* 显著下调。

1.2.4 鉴定肿瘤相关模块和其中的关键基因

在 DC 和 DE 分析之后，我们发现显著 DC 模块内的基因相比非 DC 模块的基因具有更高的差异表达打分 (*DES*)，秩和检验表明两类模块内基因的差异表达打分具有显著差异 (图 3)。这一现象表明在肿瘤中得到或失去相关性的 DC 基因也倾向于发生表达水平的变化，也佐证了 DC 和 DE 基因均与失调的生物学过程有关[2,6-9,32-33]。基于 DC 和 DE 的正相关关联，我们将同时满足 DC 和 DE 的基因模块定义为肿瘤相关模块 (图 1 底部)。肿瘤相关模块又进一步分为激活模块 (显著 DC 且显著上调) 和失能模块 (显著 DC 且显著下调)。在激活模块中，当一个基因是 DC 关键基因且其合并差异表达 *P* 值 ($P_{combined}$)<0.05 时，该基因为激活模块关键基因；在失能模块中，当一个基因是 DC 关键基因且其合并差异表达 *P* 值 ($P_{combined}$)>0.95 时，该基因为失能模块的关键基因。

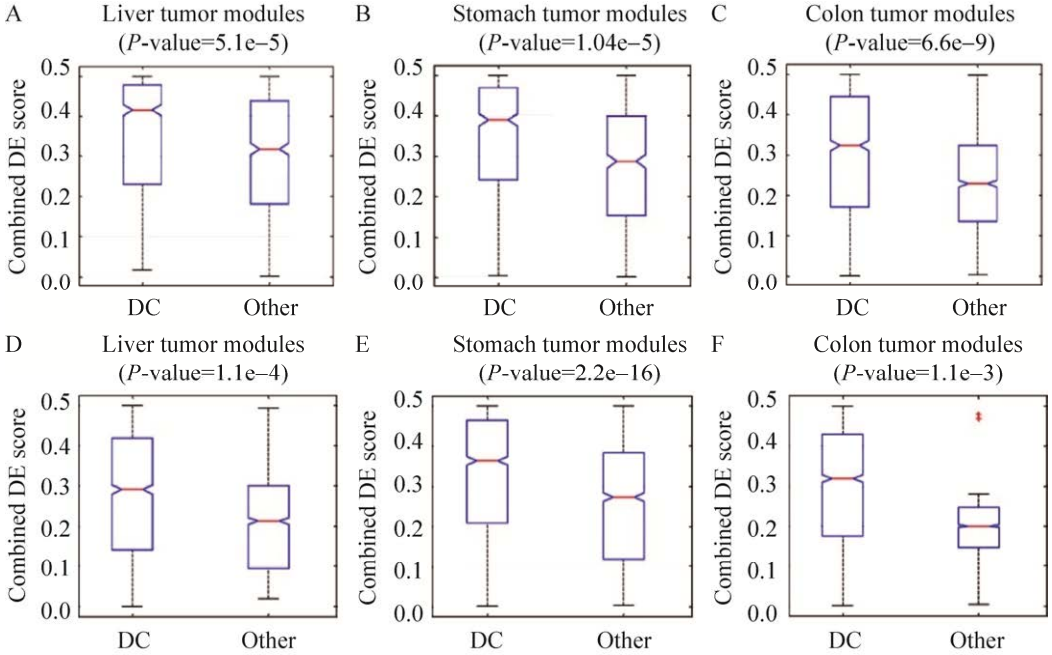

图 3 DC 模块基因与非 DC 模块基因的相关性比较

Fig. 3 Correlation between DC and DE genes. Tumor (A–C) and normal (D–F) co-expression modules of liver, stomach and colon cancer were respectively divided into significantly DC modules and the other ones. The combined DE scores of genes included in significantly DC modules are compared with those of the other modules through Wilcoxon rank-sum test and shown in box plots. The Wilcoxon rank-sum test *P*-values are shown above the box plots.

1.2.5 利用随机森林评估关键基因

在得到肝癌、胃癌和结直肠癌各自的肿瘤相关模块和其中的关键基因后，我们将上述 3 种癌症的各两组微阵列数据作为训练集，将关键基因的表达量作为特征训练了针对 3 种癌症的随机森林模型。我们利用这些模型分析一系列测试集，以考察其区分肿瘤和癌旁样本的效能。测试集包括从 GEO 数据库中下载的与肝癌、胃癌和结直肠癌相关的 3 组微阵列数据 GSE45436 (41 个癌旁和 62 个肿瘤样品)、GSE7993 (10 个癌旁和 10 个肿瘤样品) 和 GSE18105 (17 个癌旁和 17 个肿瘤样品)，以及 TCGA 数据库下载的与肝癌 (50 例肿瘤和癌旁组织)、胃癌 (27 例肿瘤和癌旁组织) 和结直肠癌 (41 例肿瘤和癌旁组织) 相关的转录组测序 (RNA Sequencing) 数据。

1.2.6 关键基因的 Kaplan–Meier 生存分析

利用文本挖掘的方式，我们搜索了每种癌症的关键基因是否已被报道为癌症预后标志物。针对未报道的关键基因，我们进一步利用 Kaplan-Meier 生存曲线分析其对癌症存活率的影响。生存曲线使用了来自 TCGA 数据库的 3 种癌症的临床数据和 RNA 测序数据，分别包括 363、348 和 430 位肝癌、胃癌和结直肠癌患者。具体方法为以特定基因在患者中的表达量从大到小对患者排序，前 1/4 的患者为高表达组，后 1/4 的患者为低表达组，基于总体生存率 (Overall survival rate) 绘制两类患者的生存曲线，当 Log-rank 检验 *P* 值小于 0.05 时认为两类患者的预后存在显著差异，该基因为癌症预后标志物。

1.2.7 肿瘤相关模块基因的通路富集分析

我们使用 R 语言中程序包“ReactomePA”和“clusterProfiler”分别对激活和失能模块基因进行通路富集分析。两个软件包均使用双侧费舍尔精确检验，分别基于 Reactome[34]和 KEGG 数据库识别显著通路，费舍尔精确检验 *P* 值小于 0.05 的通路具有统计学意义。

2 结果与分析

2.1 整合 DC 和 DE 的分析框架 DC&DEmodule

为了有效研究共表达模块的 DC 和 DE 特征，我们开发了一种新颖的整合 DC 和 DE 特征的分析框架 DC&DEmodule 并将其应用于肝癌、胃癌和结直肠癌的 6 个微阵列表达谱中，最终分别鉴定了 2 个 (122 个基因)、5 个 (108 个基因) 和 2 个 (23 个基因) 激活模块，以及 5 个 (144 个基因)、5 个 (281 个基因) 和 1 个 (68 个基因) 失能模块。详见补充表 S3–S5。

为了考察肿瘤相关模块涉及的生物过程，我们分别基于 KEGG 和 Reactome 数据库进行了通路富集分析。针对肝癌、胃癌和结肠癌的激活模块，我们分别鉴定到 21、44 和 27 条 KEGG 通路以及 187、418 和 214 条 Reactome 通路。针对肝癌、胃癌和结肠癌的失能模块，我们鉴定到 62、49 和 72 条 KEGG 通路以及 151、153 和 243 条 Reactome 通路 (详见补充表 S6)。

我们首先考察了不同癌症共有和差异的通路 (详见补充表 S7)。我们发现 22 条 KEGG 通路和 249 条 Reactome 通路在至少两种癌症的激活模块中富集 (图 4A 和 4C)。这些通路大多涉及细胞生长和凋亡，例如“DNA 复制”，“细胞周期”和“p53 信号传导通路”以及一些癌症相关的代谢过程，例如“嘧啶代谢”和“糖酵解/糖异生”等 (图 4E 和 4G)。这些结果与目前学术界的认知是一致的，即所有癌症都呈现不受控制的细胞增殖以及核苷酸生物合成的上调[35]。其中，我们发现了一条 Reactome 通路“SLIT 和 ROBO1 的表达调控 (Regulation of expression of SLITs and ROBOs)”在 3 种癌症中均被富集到。先前也有研究表明 SLIT2 和 ROBO1 蛋白在多种癌症 (例如乳腺癌、前列腺癌、卵巢癌、肺癌、肝癌和结肠癌) 中显著差异且被认为是多种致癌信号的重要调控因子[36-38]。在 3 种癌症的失能模块中，我们发现 41 条 KEGG 通路和 89 条 Reactome 通路在至少两种癌症中被富集

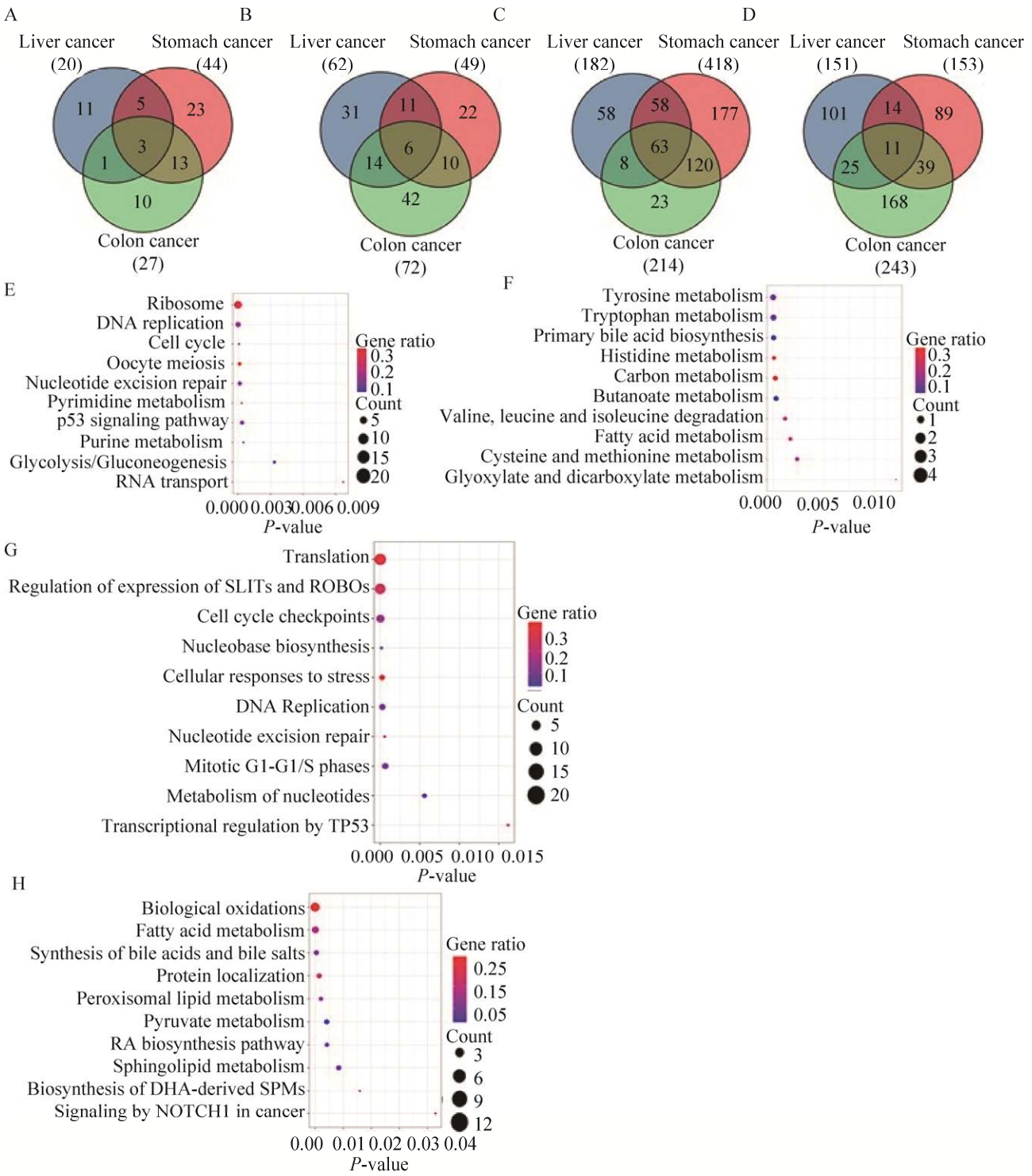

图 4　通路富集分析总览

Fig. 4　Overview of pathway enrichment analysis. Venn diagrams of activated (A) and dysfunctional (B) KEGG pathways in liver, gastric and colon cancer. Venn diagrams of activated (C) and dysfunctional (D) Reactome pathways in liver, gastric and colon cancer. (E) Ten representative activated KEGG pathways shared by at least two cancers. (F) Ten representative dysfunctional KEGG pathways shared by at least two cancers. (G) Ten representative activated Reactome pathways shared by at least two cancers. (H) Ten representative dysfunctional Reactome pathways shared by at least two cancers.

(图 4B 和 4D)。这些通路多与代谢过程有关 (图 4F 和 4H),包括氨基酸代谢 (例如“氨基酸的生物合成” “组氨酸代谢” “色氨酸代谢”和“酪氨酸代谢”)、脂质代谢 (例如“脂肪酸降解” “甾类激素生物合成”和“不饱和脂肪酸的生物合成”)、碳水化合物代谢 (例如“柠檬酸循环”和“丙酮酸代谢”) 以及能量代谢 (例如“硫代谢”和“氮代谢”)。也有研究表明，肿瘤中代谢相关基因表达呈总体下调趋势[35]，与我们的研究一致。此外，我们还发现了一些癌症特异性的代谢通路,例如“胆汁分泌”、“胆汁酸和胆汁盐代谢”作为肝脏特有的生物过程仅在肝癌的失能模块中富集到[39]。总体而言，通路富集分析表明我们鉴定的肿瘤相关模块与关键的癌症通路息息相关。

2.2 方法学比较

为了评估 DC&DEmodule 框架的性能，我们将 DC&DEmodule 与现有的几种预测肿瘤相关通路的方法进行了比较。这些方法包括独立的 DC 分析 (GSCA)、独立的 DE 分析 (Student’s *t*-test) 以及现有整合 DC 和 DE 分析的方法 (DECODE)。我们利用以上方法分别处理 3 种癌症的微阵列表达谱，并对它们的结果进行 KEGG 和 Reactome 通路富集分析 (详见补充表 S8)。结果如下：独立的差异共表达分析方法即 GSCA 富集了 37、42 和 54 条被报道过的分别与 3 种癌症相关的 KEGG 通路和 47、52 和 91 条被报道过的分别与 3 种癌症相关的 Reactome 通路；独立的差异表达分析方法即 T 检验富集了 17、15 和 28 条被报道过的分别与三种癌症相关的 KEGG 通路以及 100、112 和 30 条被报道过的分别与 3 种癌症相关的 Reactome 通路；DECODE 富集了 41、48 和 49 条被报道过的分别与 3 种癌症相关的 KEGG 通路以及 97、117 和 118 条被报道过的分别与 3 种癌症相关的 Reactome 通路；DC&DEmodule 富集了 83、93 和 99 条被报道过的分别与 3 种癌症相关的 KEGG 通路以及 338、571 和 457 条被报道过的分别与 3 种癌症相关的 Reactome 通路 (结果 2.1 部分)。

为了比较不同方法在通路预测上的效能，根据 KEGG 和 Reactome 数据库的通路注释，我们收集到与肝癌、胃癌和结直肠癌有关的 58、57 和 57 条 KEGG 通路以及 190、187 和 188 条 Reactome 通路 (详见补充表 S9)。将上述各类方法的通路富集结果与已知的癌症相关通路进行比较后发现，DC&DEmodule 在鉴定通路总数及已知的癌症相关通路方面均具有最高的灵敏度 (图 5A 和 5B)。

2.3 关键基因分析

为了进一步探究模块内基因的功能，我们分别从肝癌、胃癌和结肠癌的激活模块中筛选了 17、69 和 11 个关键基因 (详见材料与方法)。随后基于这些关键基因训练了 3 种癌症的随机森林分类器并对 GEO 和 TCGA 数据库中多组测试集的肿瘤和癌旁样本进行了分类 (数据下载地址见补充表 S10)。图 6A–C 展示了各组数据分类结果的 ROC 曲线，曲线下的面积分别为 0.99、0.91、0.96、0.90、0.88 和 0.95，其平均准确性达到 93%，证明了关键基因与肿瘤和癌旁之间的差异密切相关，可以有效分离两者并且受不同来源数据的批次处理影响较小。

为了研究关键基因与肿瘤预后的关系，我们对关键基因进行了生存分析。首先对鉴定到的 17、69 和 11 个关键基因进行文本挖掘。最终得到 16、44 和 4 个已报道与 3 种癌症相关的关键基因，其中包括 15、35 和 3 个已知的预后标志物 (详见补充表 S11) 。对于未证明的关键基因，我们进行了 Kaplan-Meier 生存分析，发现了 3 个新的癌症预后生物标志物 TCEB1、RFC4 和 TRPC4AP。第一，有学者发现 TCEB1 高表达时乳腺癌患者的预后较差[40]，而我们发现 TCEB1 的高表达同样会导致肝癌患者的存活率降低 (图 6D)。第二，以往研究发现 RFC4 高表达时宫颈癌患者的预后良好[41]但结肠癌和肝癌患者的预后较差[42-43]。在本研究中我们发现 RFC4 高表达提高了胃癌患者的生存率 (图 6E)。第三，根据“人类蛋白质图谱 (Human protein atlas)”，TRPC4AP 高表达时肝癌和肾癌患者预后较差[44]，而在这里我们发现 TRPC4AP 高表达提高了结直肠癌患者的生存率 (图 6F)。经调

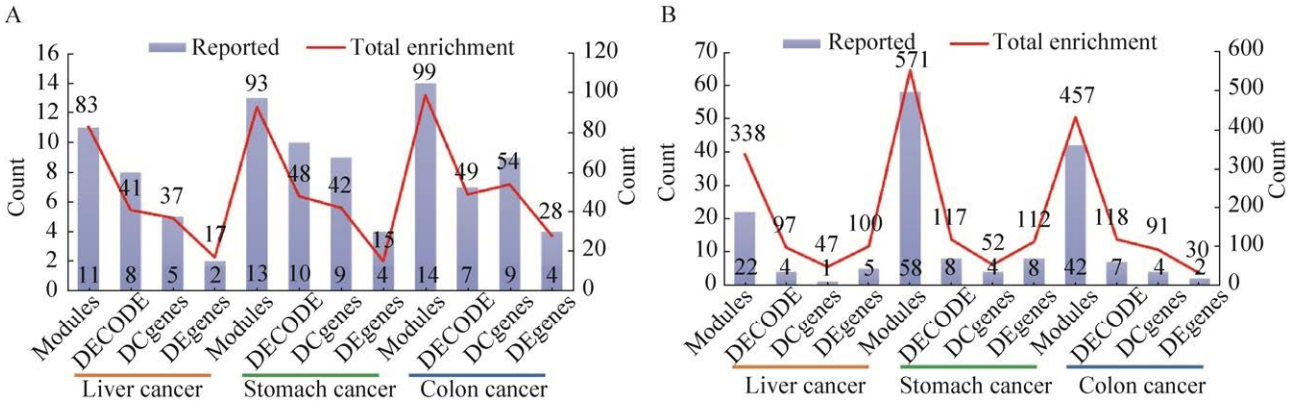

图 5　通路富集分析结果的比较

Fig. 5　Comparison of the pathway enrichment analysis. (A) The number of total KEGG pathways (line charts) and those reportedly cancer-related (bar charts) identified by different methods (DC–DE integration analytical framework, independent DC analysis, independent DE analysis, and DECODE) in liver, gastric and colon cancer. (B) The number of total Reactome pathways (line charts) and those reportedly cancer-related (bar charts) identified by different methods.

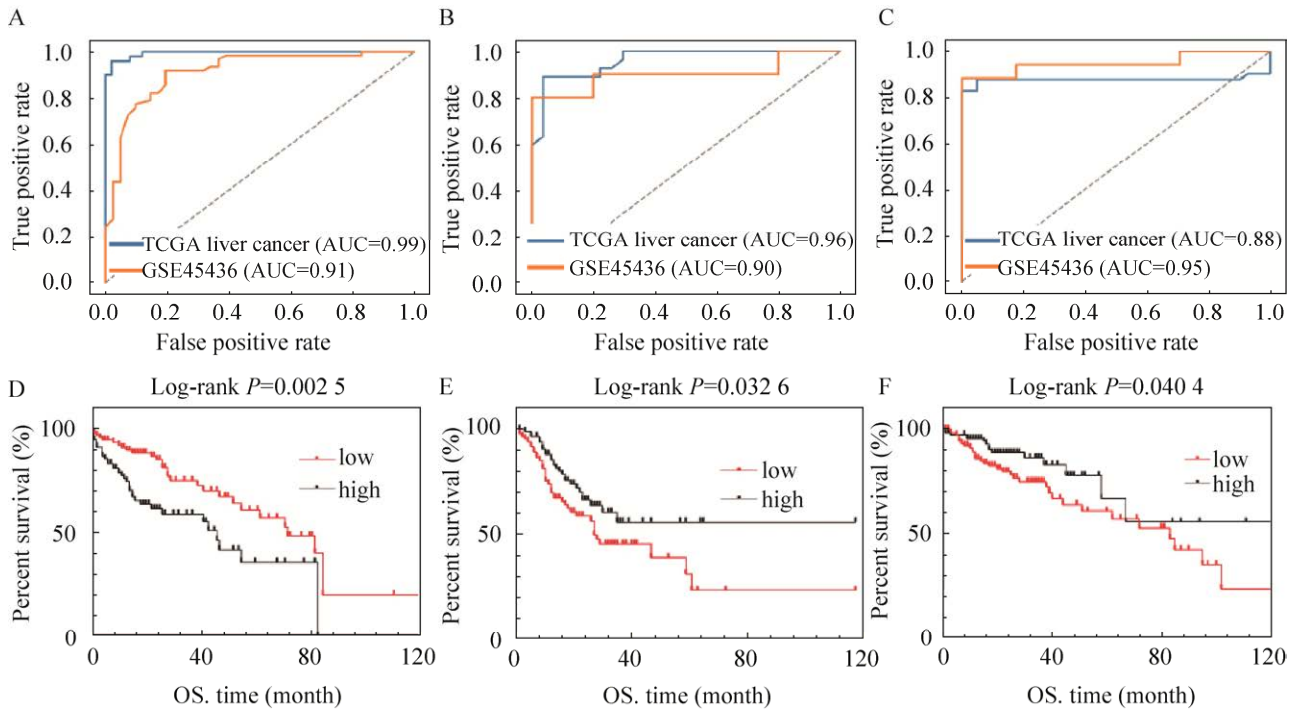

图 6　肿瘤相关模块和关键基因的评估

Fig. 6　Evaluation of the tumor-associated modules and hub genes. ROC curves of using the random forest classifiers trained by the liver (A), gastric (B) and colon (C) activated hub genes to classifier tumor and normal samples in additional datasets. (D) Overall survival plots of liver cancer patients in TCGA stratified by TCEB1 expression (log-rank *P*=0.002 5). (E) Overall survival plots of gastric cancer patients in TCGA stratified by RFC4 expression (log-rank *P* =0.032 6). (F) Overall survival plots of colon cancer patients in TCGA stratified by TRPC4AP expression (log-rank *P*=0.040 4).

研，3 个新标志物均在癌症中具有潜在的调控作用。例如，TCEB1 编码蛋白质 Elongin C，它充当 E3 泛素连接酶复合体，介导 HIF-1α 降解，从而影响癌症中的血管生成和细胞增殖[45]。RFC4 充当 DNA 聚合酶的引物识别因子，涉及“DNA 复制”和“错配修复”通路[46]。TRPC4AP 充当 Myc 特异性受体，通过将 Myc 桥接至 E3 连接酶复合物来促进 Myc 泛素化，Myc 的降解影响细胞周期和肿瘤进展[47]。同一基因 (例如 RFC4 和 TRPC4AP) 往往在不同的癌症中具有相反的作用。与我们的发现一致，Li 等发现 RFC4 可以根据肿瘤的细胞和组织学特征充当原癌基因或抑癌基因[48]。综上，癌症预后是个复杂多样的过程，所以对基因的精确调控机制还需要更严格的实验和评估。

3 讨论

当前，尽管蛋白质质谱和单细胞测序等新兴的实验技术正在迅速发展，但传统技术产出的大量微阵列数据和 RNA 测序数据尚未得到充分挖掘。因此，亟需开发有效的数据分析方法对公共数据库中大批量转录组表达谱进行整合和二次分析。此外，基因间的相互作用关系复杂而多样，通过鉴定基因共表达模块来研究生物系统中基因间的关系，可以在分子水平上发现影响癌症和其他疾病发生发展的关键因素。在本研究中，我们提出了一种整合 DC 分析和 DE 分析同时整合不同来源数据集的分析框架 DC&DEmodule，可以从疾病/健康表达谱中识别激活和功能失调的共表达基因模块。一系列评估和比较表明，DC&DEmodule 框架可以有效发现肝癌、胃癌和结直肠癌中激活和失调的生物学通路和关键基因。

此外，针对 3 种癌症的关键基因的生存分析表明，同一基因在不同的癌症类型的预后中往往具有相反的作用。因此，本研究中整合同种癌症类型的多组数据集将有助于消除批次效应并帮助阐明不同癌症类型之间的真正差异。总而言之，我们相信新的分析框架可以整合公共数据库中快速累积的表达谱，帮助研究人员发现更多人类疾病中功能失调的生物过程和关键基因，为揭露复杂疾病的发生发展机制提供新的视角。

REFERENCES

[1] Liu ZP, Wang Y, Zhang XS, et al. Network-based analysis of complex diseases. IET Syst Biol, 2012, 6(1): 22-33.

[2] Stuart JM, Segal E, Koller D, et al. A gene-coexpression network for global discovery of conserved genetic modules. Science, 2003, 302(5643): 249-255.

[3] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics, 2008, 9(1): 1-13.

[4] Oldham MC, Konopka G, Iwamoto K, et al. Functional organization of the transcriptome in human brain. Nature Neurosci, 2008, 11(11): 1271-1282.

[5] Dewey FE, Perez MV, Wheeler MT, et al. Gene coexpression network topology of cardiac development, hypertrophy, and failure. Circ Cardiovasc Genet, 2011, 4(1): 26-35.

[6] Mitra K, Carvunis AR, Ramesh SK, et al. Integrative approaches for finding modular structure in biological networks. Nat Rev Genet, 2013, 14(10): 719-732.

[7] Segal E, Shapira M, Regev A, et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet, 2003, 34(2): 166-176.

[8] Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet, 2011, 12(1): 56-68.

[9] Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet, 2004, 5(2): 101-113.

[10] de la Fuente A. From ‘differential expression’ to ‘differential networking’–identification of dysfunctional regulatory networks in diseases. Trends Genet, 2010, 26(7): 326-333.

[11] Kostka D, Spang R. Finding disease specific alterations in the co-expression of genes. Bioinformatics, 2004, 20(suppl_1): i194-i199.

[12] Carter SL, Brechbühler CM, Griffin M, et al. Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics, 2004, 20(14): 2242-2250.

[13] Hu R, Qiu X, Glazko G, et al. Detecting intergene correlation changes in microarray analysis: a new approach to gene selection. BMC Bioinformatics, 2009, 10(1): 1-9.

[14] Mentzen WI, Floris M, de la Fuente A. Dissecting the dynamics of dysregulation of cellular processes in mouse mammary gland tumor. BMC Genomics, 2009, 10(1): 1-12.

[15] Choi YJ, Kendziorski C. Statistical methods for gene set co-expression analysis. Bioinformatics, 2009, 25(21): 2780-2786.

[16] He D, Liu ZP, Honda M, et al. Coexpression network analysis in chronic hepatitis B and C hepatic lesions reveals distinct patterns of disease progression to hepatocellular carcinoma. J Mol Cell Biol, 2012, 4(3): 140-152.

[17] Drag M, Skinkyté-Juskiené R, Do DN, et al. Differential expression and co-expression gene networks reveal candidate biomarkers of boar taint in non-castrated pigs. Sci Rep, 2017, 7(1): 1-18.

[18] Lui TW, Tsui NB, Chan LW, et al. DECODE: an integrated differential co-expression and differential expression analysis of gene expression data. BMC Bioinformatics, 2015, 16(1): 1-15.

[19] Mah WC, Thurnherr T, Chow PK, et al. Methylation profiles reveal distinct subgroup of hepatocellular carcinoma patients with poor prognosis. PLoS ONE, 2014, 9(8): e104158.

[20] Yildiz G, Arslan-Ergul A, Bagislar S, et al. Genome-wide transcriptional reorganization associated with senescence-to-immortality switch during human hepatocellular carcinogenesis. PLoS ONE, 2013, 8(5): e64016.

[21] D’Errico M, de Rinaldis E, Blasi MF, et al. Genome-wide expression profile of sporadic gastric cancers with microsatellite instability. Eur J Cancer, 2009, 45(3): 461-469.

[22] Wang Q, Wen YG, Li DP, et al. Upregulated INHBA expression is associated with poor survival in gastric cancer. Med Oncol, 2012, 29(1): 77-83.

[23] Uddin S, Ahmed M, Hussain A, et al. Genome-wide expression analysis of Middle Eastern colorectal cancer reveals FOXM1 as a novel target for cancer therapy. Am J Pathol, 2011, 178(2): 537-547.

[24] Lin G, He X, Ji H, et al. Reproducibility Probability Score—incorporating measurement variability across laboratories for gene selection. Nat Biotechnol, 2006, 24(12): 1476-1477.

[25] Eschrich SA, Hoerter AM, Bloom GC, et al. Tissue-specific RMA models to incrementally normalize Affymetrix GeneChip data. Annu Int Conf IEEE Eng Med Biol Soc, 2008: 2419-2422.

[26] Gaudet P, Argoud-Puy G, Cusin I, et al. neXtProt: organizing protein knowledge in the context of human proteome projects. J Proteome Res, 2013, 12(1): 293-298.

[27] Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for R. Bioinformatics, 2008, 24(5): 719-720.

[28] Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci USA, 2003, 100(16): 9440-9445.

[29] Darlington RB, Hayes AF. Combining independent *P* values: extensions of the Stouffer and binomial methods. Psychol Methods, 2000, 5(4): 496-515.

[30] Stouffer SA. A study of attitudes. Sci Am, 1949, 180(5): 11-15.

[31] Rhodes DR, Barrette TR, Rubin MA, et al. Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. Cancer Res, 2002, 62(15): 4427-4433.

[32] Hartwell LH, Hopfield JJ, Leibler S, et al. From

molecular to modular cell biology. Nature, 1999, 402(6761): C47-C52.

[33] Alon U. Biological networks: the tinkerer as an engineer. Science, 2003, 301(5641): 1866-1867.

[34] Croft D, O'kelly G, Wu G, et al. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res, 2010, 39(suppl_1): D691-D697.

[35] Hu J, Locasale JW, Bielas JH, et al. Heterogeneity of tumor-induced gene expression changes in the human metabolic network. Nat Biotechnol, 2013, 31(6): 522-529.

[36] Gara RK, Kumari S, Ganju A, et al. Slit/Robo pathway: a promising therapeutic target for cancer. Drug Discovery Today, 2015, 20(1): 156-164.

[37] Avci ME, Konu O, Yagci T. Quantification of SLIT-ROBO transcripts in hepatocellular carcinoma reveals two groups of genes with coordinate expression. BMC Cancer, 2008, 8(1): 1-11.

[38] Huang T, Kang W, Cheng AS L, et al. The emerging role of Slit-Robo pathway in gastric and other gastro intestinal cancers. BMC Cancer, 2015, 15(1): 1-9.

[39] Huang Q, Tan Y, Yin P, et al. Metabolic characterization of hepatocellular carcinoma using nontargeted tissue metabolomics. Cancer Res, 2013, 73(16): 4992-5002.

[40] Lindskog C. The Human Protein Atlas–an important resource for basic and clinical research. Expert Rev Proteomics, 2016, 13(7): 627-629.

[41] Tang X, Xu Y, Lu L, et al. Identification of key candidate genes and small molecule drugs in cervical cancer by bioinformatics strategy. Cancer Manag Res, 2018, 10: 3533.

[42] Xiang J, Fang L, Luo Y, et al. Levels of human replication factor C4, a clamp loader, correlate with tumor progression and predict the prognosis for colorectal cancer. J Transl Med, 2014, 12(1): 1-11.

[43] Yang WX, Pan YY, You CG. CDK1, CCNB1, CDC20, BUB1, MAD2L1, MCM3, BUB1B, MCM2, and RFC4 may be potential therapeutic targets for hepatocellular carcinoma using integrated bioinformatic analysis. BioMed Res Int, 2019: 1245072.

[44] Uhlén M, Fagerberg L, Hallström BM, et al. Tissue-based map of the human proteome. Science, 2015, 347(6220): 1260419.

[45] Kamura T. Cullin-based E3 family. Tanpakushitsu Kakusan Koso, 2006, 51(10 Suppl): 1167-1172.

[46] Ellison V, Stillman B. Biochemical characterization of DNA damage checkpoint complexes: clamp loader and clamp complexes with specificity for 5′ recessed DNA. PLoS Biol, 2003, 1(2): e33.

[47] Choi SH, Wright JB, Gerber SA, et al. Myc protein is stabilized by suppression of a novel E3 ligase complex in cancer cells. Genes Dev, 2010, 24(12): 1236-1241.

[48] Li Y, Gan S, Ren L, et al. Multifaceted regulation and functions of replication factor C family in human cancers. Am J Cancer Res, 2018, 8(8): 1343-1355.

(本文责编　陈宏宇)