Computer Science and Application 计算机科学与应用, 2020, 10(10), 1820-1833
Published Online October 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.1010192

Hans汉斯

整合多组学数据搜寻胃癌的驱动基因

杨正伟，张　娟

华北电力大学，北京
Email: 1778949216@, Zhangjuan55@ncepu.edu.cn

收稿日期：2020年10月6日；录用日期：2020年10月21日；发布日期：2020年10月28日

摘　要

目的：寻找对胃癌的发生发展有驱动作用的基因和信号通路，为胃癌的靶向治疗提供可能的靶点。方法：从TCGA下载胃癌的RNA-Seq和DNA拷贝数变异数据，利用Perl脚本和R语言中的edgR包筛选出既差异表达且拷贝数发生变异的基因。将筛选出的基因进行基因功能富集分析和通路分析。再将富集过后的基因上传String数据库，生成蛋白质相互作用网络(PPI)，进而从PPI网络中分离出结合度高的稠密子网络，最后结合MalaCards疾病数据库构建胃癌的核心网络，进而寻找因差异表达与基因拷贝数呈正相关的基因。结果：本文一共筛选出26个胃癌驱动基，以及PI3K-Akt信号通路、ECM-受体相互作用、Wnt信号通路和胃酸分泌等可能与胃癌的发生发展相关的信号通路，为胃癌的诊断和治疗提供潜在的靶点。

关键词

生物信息学，胃癌，靶向治疗，DNA拷贝数变异，PPI网络

Integrating Multiple Omics Data to Search for Driver Genes for Gastric Cancer

Zhengwei Yang, Juan Zhang

North China Electric Power University, Beijing
Email: 1778949216@, Zhangjuan55@ncepu.edu.cn

Received: Oct. 6th, 2020; accepted: Oct. 21st, 2020; published: Oct. 28th, 2020

Abstract

Aims: To find driving genes and signal pathways that cause the occurrence and development of gastric cancer, in order to provide possible targets for targeted therapy of gastric cancer. Methods: This paper downloads RNA-Seq and DNA copy number variation data of gastric cancer from TCGA, uses Perl script and the edgR package in R language to screen out genes that are differentially expressed and have copy

文章引用: 杨正伟, 张娟. 整合多组学数据搜寻胃癌的驱动基因[J]. 计算机科学与应用, 2020, 10(10): 1820-1833.
DOI: 10.12677/csa.2020.1010192

number variation; then performs gene function enrichment analysis and pathway analysis of the selected genes; after then uploads the enriched genes to the String database to generate a protein interaction network (PPI), and then separates dense sub-networks with high binding degree from the PPI network. Finally, it combines with MalaCards disease database to construct the core network of gastric cancer, and then finds genes whose differential expression is positively correlated with gene copy number. Results: This paper screened a total of 26 gastric cancer driving groups, as well as PI3K-Akt signaling pathway, ECM-receptor interaction, Wnt signaling pathway, gastric acid secretion and other signaling pathways that may be related to the occurrence and development of gastric cancer to provide potential targets for the diagnosis and treatment of gastric cancer.

Keywords

Bioinformatical, Gastric Cancer, Targeted Therapy, DNA Copy Number Variation, Protein Protein Interaction Network

1. 引言

胃癌已经成为第五个最常被诊断出的癌症和癌症死亡的第三大主要原因[1]。2018 年，胃癌新病例到百万，死亡病例大约 78.3 万例，相当于全球每 12 例死亡中至少有 1 例由胃癌导致。

依据癌组织侵袭程度，将其分为早期胃癌和进展期胃癌。由于早期胃癌无特异性改变，大多患者发现时已经处于进展期[2]。虽然放化疗、手术切除等传统的胃癌治疗方式在一定程度上提高了患者的生存期，但化学抗性以及癌症的异质性和分子复杂性，使得传统的治疗方法效果极其有限，患者的 5 年生存率较低[3] [4]。靶向治疗是在细胞分子水平上，针对已经明确的致癌位点(肿瘤细胞内部的一个蛋白分子或者一个基因片段)的治疗方式。靶向药物进入体内会特异地与致癌位点相结合，使肿瘤细胞特异性死亡，而不会波及肿瘤周围的正常组织细胞。近年来，针对遗传突变和驱动肿瘤生长和侵袭的信号通路的靶向药物的开发为胃癌的个性化诊疗提供了可能，且对于部分的中晚期胃癌症起到了较好的临床效果。寻找胃癌诊疗靶点和相关生物标志物，研究胃癌发生发展过程中的分子机制是实现靶向治疗的关键和基础。

基因芯片技术和生物信息学分析方法是目前探究疾病发生发展过程中基因调控的重要方法[5]。在以往的研究中，大多数采用单一的肿瘤基因组数据进行研究分析，进而寻找治疗肿瘤的靶点。整合多种类型的基因组数据，不仅可以从多个角度展现肿瘤分子病理的变化，还可以通过研究多类型基因数据之间的相互联系，有效地挖掘出与肿瘤相关的基因以及相关的信号通路。本文通过整合TCGA数据库中胃癌的RNA-Seq和 DNA 拷贝数变异数据，运用生物信息学的分析方法，筛选那些即发生差异表达又发生 DNA 拷贝数变异的基因，以及可能与胃癌发生发展过程相关的信号通路，为寻找治疗胃癌潜在的靶点提供新的思路。

2. 材料和方法

2.1. 数据下载

癌症和肿瘤基因图谱(TCGA)计划是在 2005 年由美国国家癌症研究所和国家人类基因组研究所联合发起的一项旨在对人类全部肿瘤进行基因组测序，绘制基因变异图谱的研究计划，现已收录了 34 种肿瘤的 11,315 份组织标本多组学数据。本文的数据来源于 TCGA 官网(https://portal.gdc.cancer.gov)，从中获取

了 32 例正常组织样本和 375 例胃癌组织样本的基因表达数据(RNA-Seq)，以及所有胃癌的 DNA 拷贝数变异数据(Copy Number Variation)，其中包含 464 例正常样本，442 例患癌样本。

2.2. 数据预处理

从 Genecode 数据库(https://www.gencodegenes.org/)下载了染色体上基因的注释 GTF 文件，运用 Perl 脚本将所获取的所有 CNV 样本数据进行基因注释，表 1 是其中一例样本注释上基因信息以后的。

Table 1. The gene annotation information of CNV samples
表 1. CNV 样本的基因注释信息

GDC_Aliquot	Chromosome	Start	End	Num_Probes	Segment_Mean	Gene
5201fb7d-ec3e-4905-92c5-571073285ab7	1	75589078	75594274	5	0.6533	SLC44A5
5201fb7d-ec3e-4905-92c5-571073285ab7	1	75774375	75775015	3	−1.1047	ACADM
5201fb7d-ec3e-4905-92c5-571073285ab7	1	109684695	109697556	15	−0.9718	GSTM2
5201fb7d-ec3e-4905-92c5-571073285ab7	1	109684695	109697556	15	−0.9718	GSTM1
5201fb7d-ec3e-4905-92c5-571073285ab7	1	143759680	143942271	12	−0.5761	HIST2H3PS2
5201fb7d-ec3e-4905-92c5-571073285ab7	1	196758720	196802150	43	0.4134	CFHR3
5201fb7d-ec3e-4905-92c5-571073285ab7	1	248585893	248633912	32	−0.6809	OR2T10
5201fb7d-ec3e-4905-92c5-571073285ab7	1	248585893	248633912	32	−0.6809	OR2T11
5201fb7d-ec3e-4905-92c5-571073285ab7	2	86725947	86726731	2	−1.8311	RMND5A
5201fb7d-ec3e-4905-92c5-571073285ab7	2	97199600	97211436	14	−0.6583	ANKRD36
5201fb7d-ec3e-4905-92c5-571073285ab7	2	97524284	97545726	9	1.2656	ANKRD36B
5201fb7d-ec3e-4905-92c5-571073285ab7	2	160221312	160222900	3	−1.5077	ITGB6

注：Segment Mean 的值为 $\log 2^{\frac{copy\_number}{2}}$。

再把所有注释上基因信息后的样本数据文件用 perl 脚本进行合并与转化，在表 1 中当 $2^{1+\text{SegmentMean}}=1$ 时，定义基因的表达值为 0，即没有发生拷贝数的变异；$2^{1+\text{SegmentMean}}<0.5$ 时，定义基因的表达值为−2；$2^{1+\text{SegmentMean}}<1.5$，定义基因的表达值为−1，$2^{1+\text{SegmentMean}}>2.5$，定义基因的表达值为 1，$2^{1+\text{SegmentMean}}>3.5$，定义基因的表达值为 2，得到 cnv 的表达矩阵，如表 2 所示，这里的 1 表示基因多复制了一个，−1 则表示基因少复制了一个，2 表示基因多复制了两个或者两个以上，−2 则表示基因少复制了两个或者多个，0 则表示该基因复制正常，没有发生拷贝数变异。

Table 2. The cnv expression matrix
表 2. cnv 表达矩阵

id	TCGA-CG-4449-10A-01D-1155-01	TCGA-BR-6852-11A-01D-1881-01	TCGA-BR-7715-10A-01D-2052-01	…
NABP2	0	0	0	…
RMND5A	0	−1	0	…
PABPC1L	0	0	0	…
CLSPN	0	0	0	…
OTOA	0	0	0	…
KRTAP2-4	0	0	0	…
GOLGB1	0	0	0	…
ZNF705A	0	0	0	…
INO80B-WBP1	0	0	0	…
…	…	…	…	…

将下载的所有 RNA-Seq 数据通过 perl 脚本处理，得到 mRNA 的表达矩阵，并从 Genesymbol 数据库下载基因的注释文件，将基因对应的 id 转换为基因名，得到基因的表达矩阵，如表 3 所示。

Table 3. The gene expression matrix
表 3. 基因表达矩阵

id	TCGA-HU-A4GC-11A-11R-A251-31	TCGA-BR-6457-11A-01R-1802-13	TCGA-HU-A4GP-11A-21R-A251-31	TCGA-CG-5722-11A-02R-1602-13	…
AC127070.3	1	0	1	1	…
RN7SL680P	1	1	1	0	…
B3GALNT1P1	9	1	18	1	…
LINC02099	0	1	1	0	…
AC006065.5	0	0	0	0	…
ACKR2	127	65	191	70	…
OR7A3P	0	0	0	0	…
CXXC1P1	0	0	0	0	…
CEACAM19	300	41	534	58	…
…	…	…	…	…	…

2.3. 基因的筛选

将得到的基因表达矩阵导入 R，运用 Bioconductor 的软件包 edgeR [6]，以 $p < 0.05$ 和$|\log FC| > 1.5$ 作为筛选差异基因的标准，筛选出胃癌正常组织和癌症组织之间的差异表达的基因，得到差异表达基因矩阵。对 cnv 的表达矩阵运用 Perl 脚本进行皮尔逊卡方检验，将检验得到的 Pvalue 值进行校正得到 adjPvalue (校正后的 Pvalue)，以 adjPvalue < 0.05 为筛选标准，筛选出发生拷贝数变异的差异基因，得到发生 DNA 拷贝数变异的差异表达矩阵。筛选出既发生差异表达又发生 DNA 拷贝数变异的基因。

2.4. GO 富集分析和 KEGG 通路分析

生物学过程通常由一组基因组成，而不是单独的个别基因。富集分析的主要基础是，如果在给定的研究中生物学过程异常，则通过高通量筛选技术将共同发挥作用的基因，即具有较高的富集潜力的基因选作相关群体。这样的理由可以使大型基因列表的分析从单个面向基因的视图转移到基于相关基因组的分析。因为分析结论是基于一组相关基因而不是单个基因，所以它增加了研究者识别与所研究的生物学现象最相关的正确生物学过程的可能性[7]。

DAVID 是由美国科学家建立的注释和富集分析数据库，用于基因列表或大规模蛋白质的生物特征数据描述，旨在为从基因组研究衍生的大量基因提供功能解释，为研究人员提供一套全面的功能注释工具[8]。为了了解筛选出来的差异表达基因在胃癌的发生、发展过程中所起的生物学作用，需要给予这些基因详细的生物学注释。将合并差异表达基因后得到的上、下调基因分别上传到 DAVID 在线数据库进行 GO 富集分析和 KEGG 通路富集分析，以 Pvalue < 0.05 为统计截断标准，筛选出可能参与胃癌发生发展过程的关键通路。

2.5. PPI 网络及重要子网络

String 数据库(http://www.string-db)是一款可以预测并评估蛋白质间相互作用的在线数据库，库旨在收集，评分和整合所有可公开获得的蛋白质间相互作用信息的来源，并通过计算预测对这些来源进行补

充[9]。为了进一步了解所筛选出来的差异表达基因之间的相互作用关系，将富集分析与通路分析所得到的差异表达基因导入 String 数据库，以 0.7 的高置信阈值来构建 PPI 网络。将结果导出到 Cytoscape 软件进行可视化，并使用其中的 MCODE 插件寻找到 PPI 网络中最重要的两个子网络。结合 MalaCards 疾病数据库，找出子网络中与胃癌相关的基因，运用 Cytoscape 的插件 CytoHubba 构建与这些基因相互作用最大的基因的子网络。

3. 统计结果分析

3.1. 差异表达基因

通过 R 软件的 edgeR 包和 Perl 脚本，得到 1867 个既发生差异表达又发生 DNA 拷贝数变异的差异表达基因，包括 946 个上调的差异表达基因，921 个下调的差异表达基因，图 1 为这些差异表达基因的火山图，红色表示上调的基因，绿色表示下调的基因，中间的黑色代表正常范围内表达的基因。

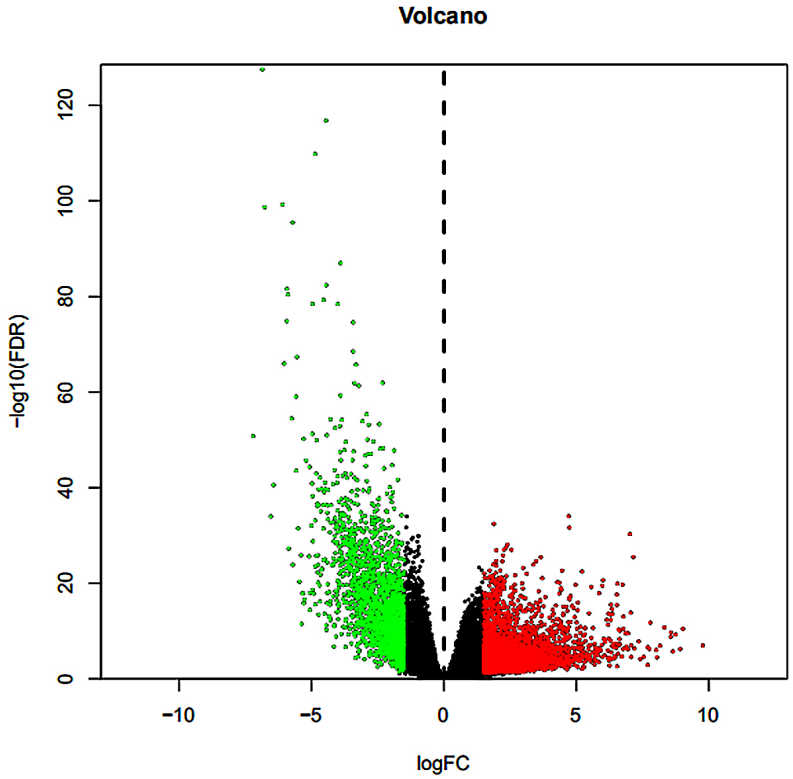

Figure 1. Volcano map of differentially expressed genes
图 1. 差异表达基因的火山图

3.2. 差异表达基因的 GO 富集分析和 KEGG 通路分析

3.2.1. 上调差异表达基因的 GO 富集分析和 KEGG 通路分析

如图 2 所示，上调差异表达基因的功能富集分析显示，上调的差异基因显著富集于趋化因子介导的信号通路、胞间信号、ERK1 和 ERK2 级联的正调控、DNA 复制、细胞分化、G1/S 有丝分裂细胞周期的转变、G 蛋白偶联受体信号通路、细胞增殖的正调控以及免疫反应等生物过程(Biological Processes)。细胞组分(Cell Component)主要聚集于细胞外区域、细胞外空间、蛋白质细胞外基质、血液微粒、胶原三聚体、血小板 α 颗粒管腔、细胞表面、细胞外基质、内质网腔和高尔基流明等。分子功能(Molecular Function)

主要与丝氨酸型内肽酶活性、序列特异性 DNA 结合、激素活性、生长因子活性、受体结合、转录激活子活性：RNA 聚合酶 II 核心启动子近端区域序列特异性结合、细胞因子活性、钙离子结合、RNA 聚合酶 II 核心启动子近端区域序列特异性 DNA 结合以及染色质结合等活动相关。

KEGG 通路分析结果显示，上调的差异基因显著富集于细胞因子与细胞因子受体的相互作用、细胞周期、蛋白质的消化吸收、脂肪的消化吸收、PI3K-Akt 信号通路、黑色素瘤、ECM-受体相互作用、胰腺分泌、膀胱癌和 Wnt 信号通路等。

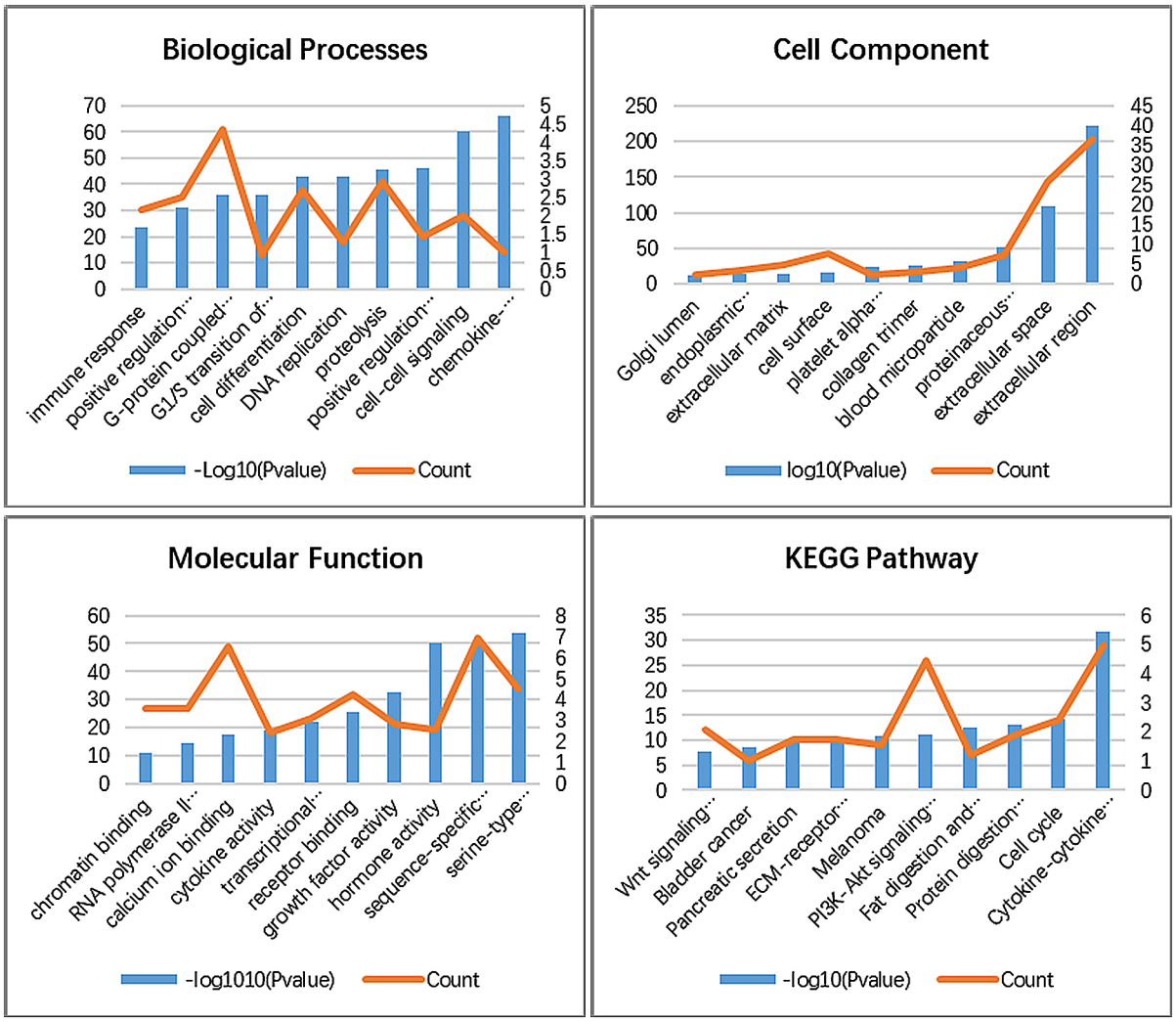

Figure 2. Functional enrichment analysis and pathway analysis of up-regulated differentially expressed genes
图 2. 上调差异表达基因的功能富集分析和通路分析

3.2.2. 下调差异表达基因的 GO 富集分析和 KEGG 通路分析

如图 3 所示，下调差异表达基因的功能富集分析显示，下调的差异基因显著富集于消化不良、离子跨膜转运、运输、钠离子迁移、细胞增殖的正调控、细胞粘附、氧化还原过程、细胞表面受体信号通路、细胞对肿瘤坏死因子的反应以及脂质代谢等生物过程(Biological Processes)。细胞组分(Cell Component)主要聚集于细胞外空间、质膜、细胞外泌体、质膜的组成部分、细胞外区域、顶质膜、细胞表面、膜的整体成分、细胞连接和受体复合物等。分子功能(Molecular Function)主要与结构分子活性、钙离子结合、脂

质结合、运输活动、氧化还原酶活性、钙调蛋白结合、肌动蛋白结合、铁离子结合、锌离子结合以及受体结合等活动相关。

KEGG 通路分析结果显示，上调的差异基因显著富集于胃酸分泌、cGMP-PKG 信号通路、cAMP 信号通路、PPAR 信号通路、化学致癌作用、脂肪的消化吸收、胰岛素分泌、胰腺分泌、蛋白质的消化吸收和 AMPK 信号通路等。

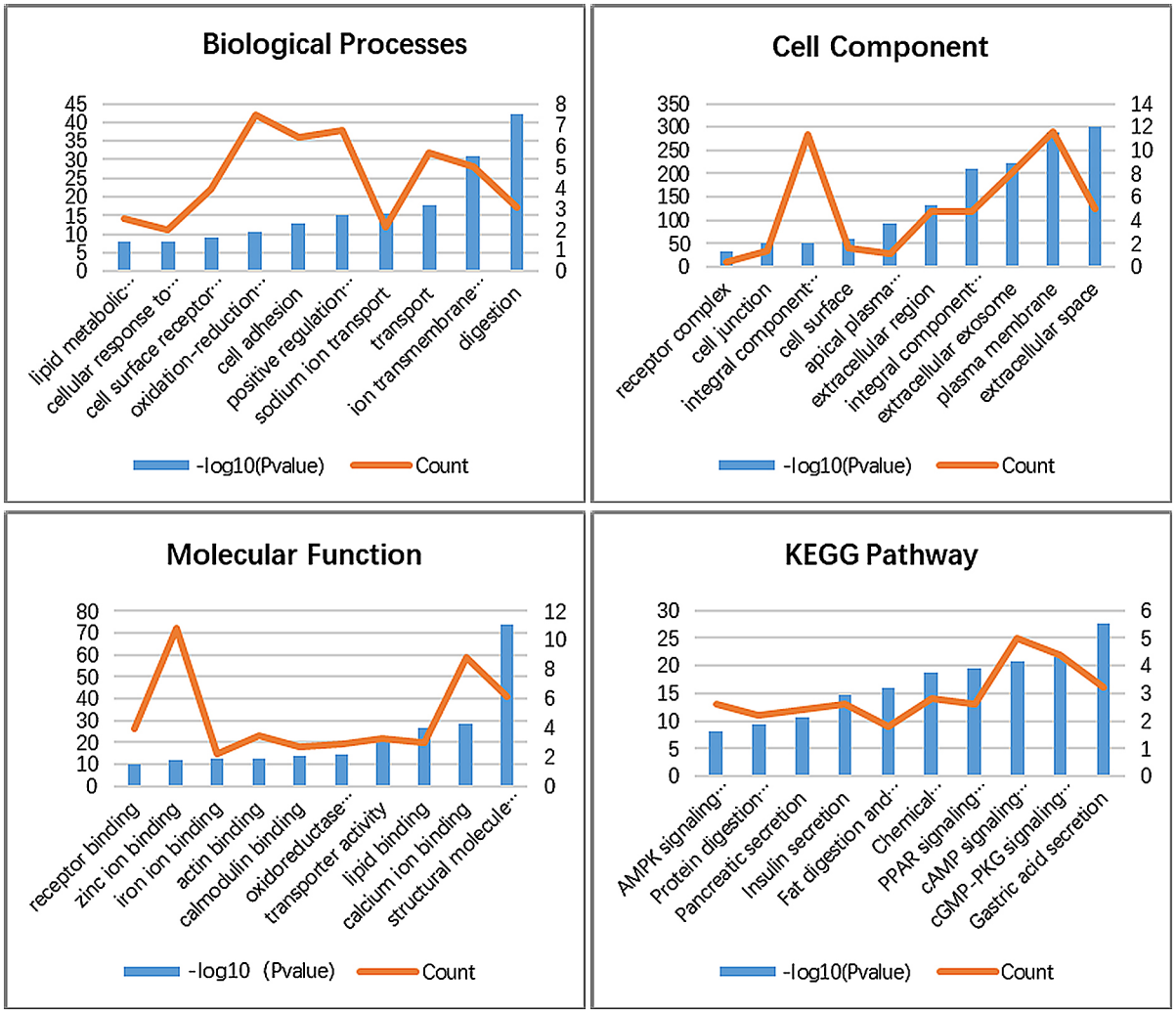

Figure 3. Functional enrichment analysis and pathway analysis of down-regulated differentially expressed genes
图 3. 下调差异表达基因的功能富集分析和通路分析

3.3. PPI 网络及重要子网络

将富集分析与通路分析所得到的 689 个上调基因与 824 个下调基因导入 String 数据库，以 0.7 的高置信阈值来构建 PPI 网络。如图 4 所示，图中的黄色表示表达上调的差异基因，蓝色表示表达下调的差异基因。将结果导入 Cytoscape 软件进行可视化，并使用其中的 MCODE 插件寻找到 PPI 网络中最重要的两个子网络。如图 5 所示，子网络 1 中包含 32 个上调基因和 39 个下调基因，子网络 2 中的 35 个基因全为上调基因。

GeneCardsSuite (http://www.genecards.org)是一个可搜索的集成数据库，该数据库自动集成了来自约

150 个网络来源的以基因为中心的数据，包括基因组，转录组学，蛋白质组学，遗传，临床和功能信息。在其中的 MalaCards 疾病数据库中找到了 499 个与胃癌相关的基因，其中有八个基因出现在子网络 1 和子网络 2 中。分别为 GAST、CCKBR、TOP2A、AURKA、CXCL8、PLK1、ECT2、TPX2。

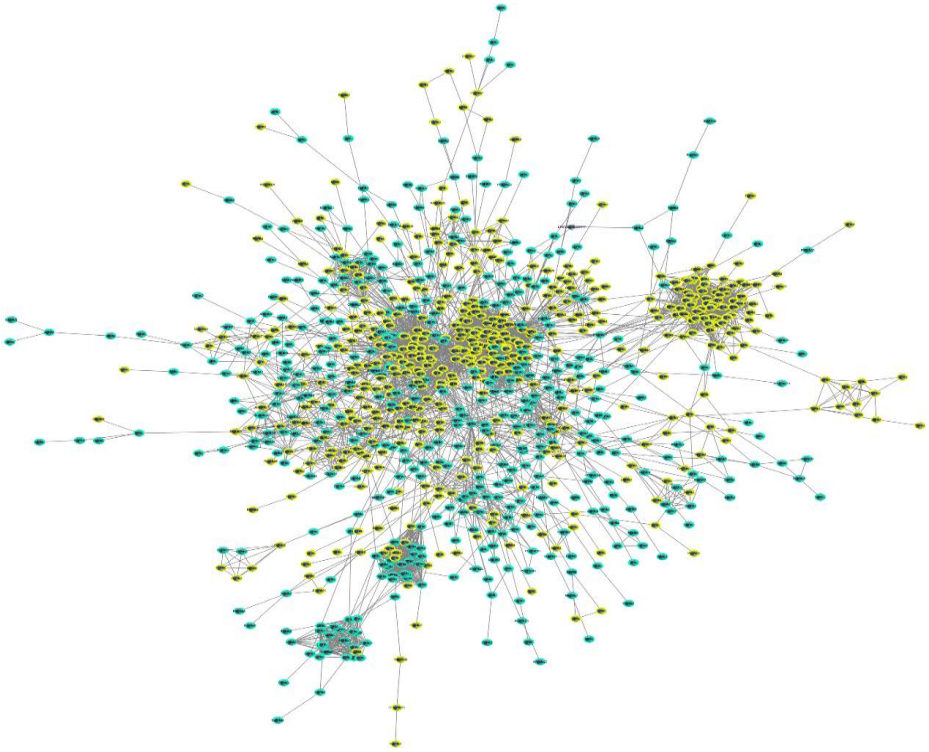

Figure 4. Protein interaction network of differentially expressed genes(PPI)
图 4. 差异表达基因的蛋白质相互作用网(PPI)

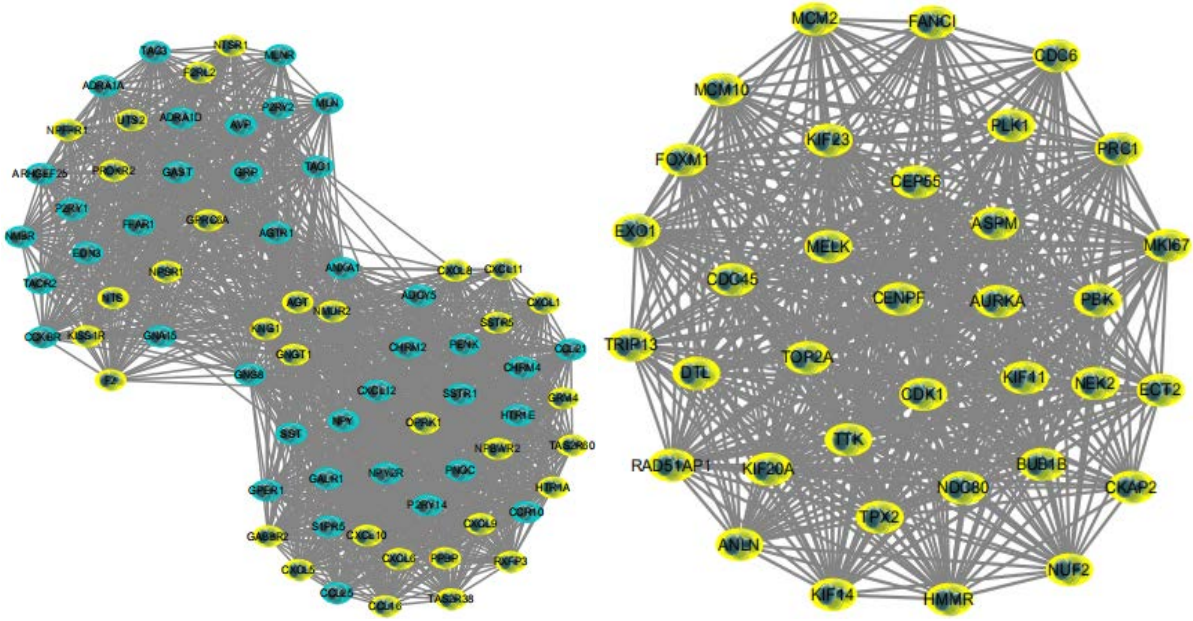

Figure 5. Two important sub-networks in PPI
图 5. PPI 中两个重要的子网络

3.4. 构建核心网络寻找胃癌驱动基因

使用 Cytoscape 的插件 CytoHubba 构建与 GAST、CCKBR、TOP2A、AURKA、CXCL8、PLK1、ECT2、TPX2 这八个基因相互作用的网络，如图 6 所示。

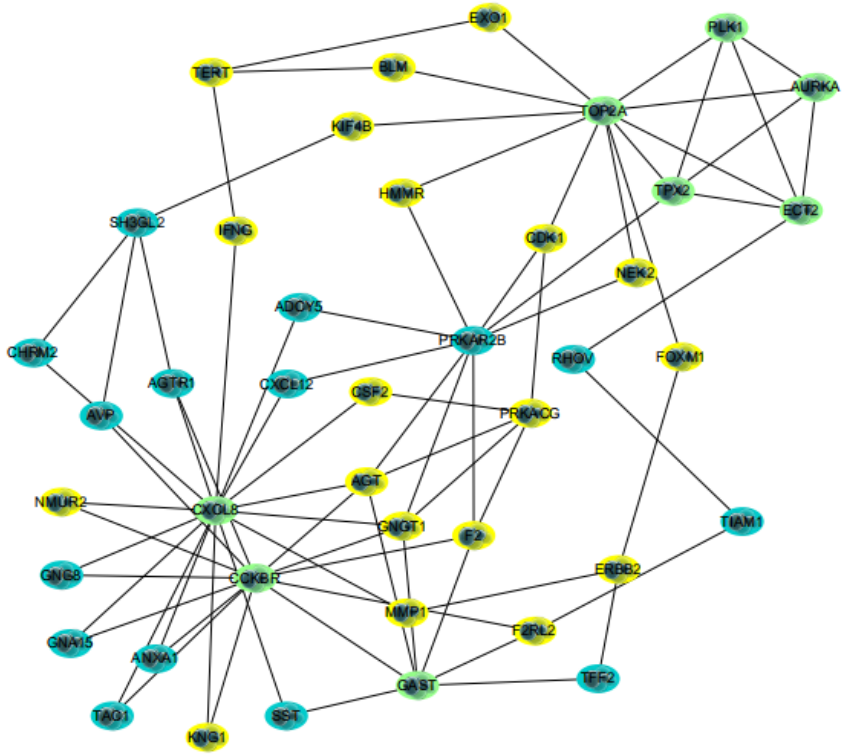

Figure 6. The core subnetwork of gastric cancer
图 6. 胃癌的核心子网络

该网络中包含上调差异基因 25 个，下调差异基因 17 个，找出其中基因差异表达与基因拷贝数呈正相关的基因，即在基因拷贝数增加的同时，基因的表达上调；相反在基因拷贝数减少时，基因的表达下调。得到如表 4 所示的 26 个驱动基因。图 7 为这 26 个基因所在的染色体位置圈图。

Table 4. Gene differential expression is positively correlated with gene copy number
表 4. 基因差异表达与基因拷贝数呈正相关的基因

Gene	copy_number	group
PLK1	1	up
AGT	1	up
BLM	1	up
FOXM1	1	up
ERBB2	2	up
NEK2	1	up
IFNG	2	up
MMP1	1	up

Continued

TERT	1	up
PRKACG	1	up
GNGT1	1	up
F2	1	up
KNG1	1	up
CDK1	1	up
EXO1	1	up
TPX2	1	up
AURKA	1	up
TOP2A	1	up
CXCL8	1	up
ECT2	1	up
SH3GL2	−1	down
TIAM1	−1	down
RHOV	−1	down
GNA15	−1	down
TFF2	−1	down
CCKBR	−1	down

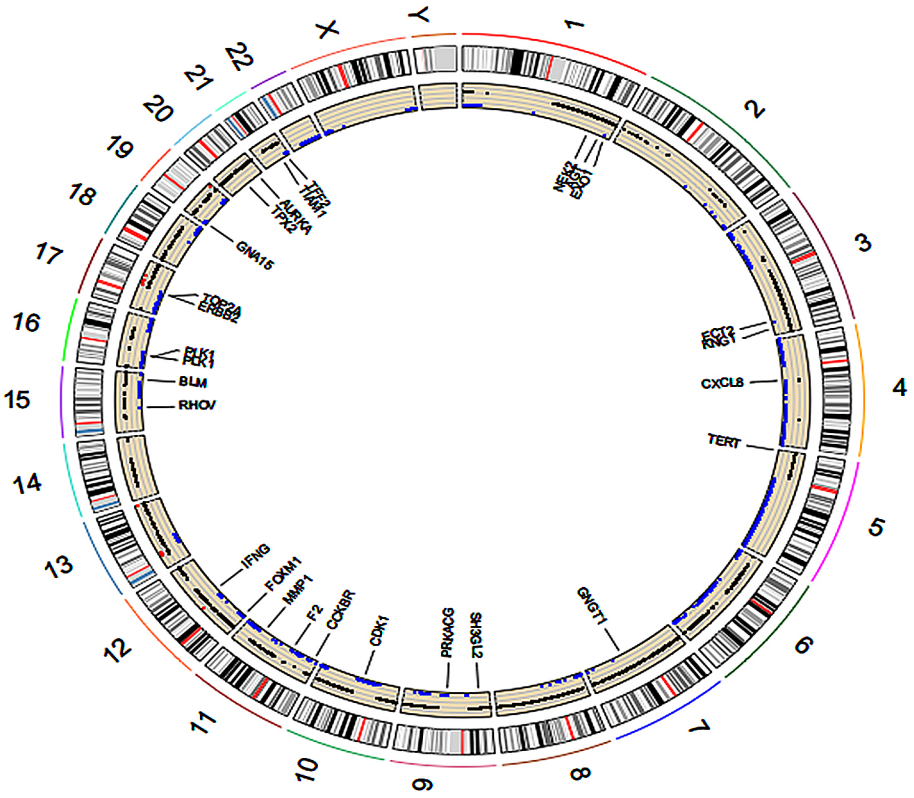

Figure 7. The location circle diagram of the driver gene on a chromosome
图 7. 驱动基因在染色体上的位置圈图

3.5. 驱动基因与胃癌的发生发展

筛选出的 26 个胃癌驱动基因中，有 19 个基因被相关的文献报道与胃癌的发生发展有关，其中极光激酶 A (AURKA)位于染色体臂 20q13 中，是一个在胃腺癌中经常扩增和过度表达的基因，Dar 等人[10]的结果表明 AURKA 调节 GSK-3β 磷酸化和活性，导致胃癌中 β-catenin/TCF 复合物的积累和活化，AURKA 过表达会增强胃癌细胞的细胞增殖。CCKBR 基因编码胃泌素和胆囊收缩素(CCK) (大脑和胃肠道的调节肽)的 G 蛋白偶联受体。该蛋白是 B 型胃泌素受体，对硫酸化和非硫酸化 CCK 类似物均具有很高的亲和力，主要存在于中枢神经系统和胃肠道中。Zhou 等[11]人的结果表明胃泌素 mRNA 的表达水平与恶性肿瘤的进展和转移密切相关，胃泌素/CCK-B 受体的自分泌或旁分泌途径的存在可能在胃癌的发病机理，特别是在进展中起重要作用。CXCL8 又名白介素(IL-8)是一种趋化因子，可吸引中性粒细胞，嗜碱性粒细胞和 T 细胞，但不吸引单核细胞。它也参与嗜中性粒细胞活化。Yasumoto 等人[12]的结果表明 IL-8 可能是由胃癌细胞局部产生的，并且 T 淋巴细胞通过嗜中性白细胞浸润直接或间接渗透到肿瘤部位，这些发现还增加了以下可能性：癌症来源的 IL-8 诱导的白细胞浸润可能在启动对肿瘤细胞的免疫反应，并可能控制肿瘤细胞的生长和转移。ECT2 基因编码的蛋白质是鸟嘌呤核苷酸交换因子和与 Rho 特异性交换因子和酵母细胞周期调节剂有关的转化蛋白。该基因的表达随 DNA 合成的开始而升高，并在 G2 和 M 期保持升高。通过与肿瘤细胞中的致癌 PARD6A-PRKCI 复合物相关联来刺激 RAC1 的活性，从而协调地驱动肿瘤细胞的增殖和侵袭。ERBB2 通常称为 HER2，HER2 扩增是胃癌以及其他各种癌症中常见的分子异常，近些年已在 20%至 30%的胃癌中检测到人类表皮生长因子受体 2 (HER2)的扩增，并与不良预后相关。Tanizaki 等人[13]研究发现在具有 HER2 扩增的胃癌细胞中，S-1 和 HER2 靶向剂的组合发挥了由 TS 抑制介导的协同抗肿瘤作用。TOP2A 基因位于 17 号染色体上 HER2 的附近，是许多化疗药物的靶标，Liang [14]等人的研究表明在胃癌中 TOP2A 基因扩增与 HER2 基因扩增显着相关。NEK2 基因编码参与有丝分裂调控的丝氨酸/苏氨酸蛋白激酶，控制有丝分裂细胞中的中心体分离和双极纺锤体形成以及减数分裂细胞中的染色质凝缩，在 G2/M 过度时调节 PLK1 活性以及 CDK 介导的磷酸化。PLK1 基因编码的 Ser/Thr 蛋白激酶属于 CDC5/Polo 亚家族。它在有丝分裂过程中高度表达，并且在许多不同类型的癌症中均发现其水平升高。癌细胞中这种蛋白质的消耗会显着抑制细胞增殖并诱导凋亡。Chen 等人[15]发现 PLK1 基因敲减与细胞周期蛋白 B 表达增加，cdc2 活性增加(但与表达水平无关)，G2/M 处的胃癌细胞蓄积，有丝分裂纺锤体形成不当，染色体分离延迟以及胞质分裂滞后或停滞有关。PLK1 表达受阻可能导致胃癌细胞有丝分裂减少甚至凋亡，揭示 PLK1 可能是胃癌的重要治疗靶点。端粒酶(TERT)是大多数真核生物中染色体末端复制必不可少的核糖核酸酶。在祖细胞和癌细胞中活跃。在正常的体细胞中无活性或极低活性端粒酶全酶复合物的催化成分，其主要活性是通过充当逆转录酶来延长端粒，该逆转录酶通过在酶的 RNA 成分内复制模板序列，在染色体末端添加简单的序列重复。Yasui 等人[16]通过逆转录聚合酶链反应在人胃癌和非肿瘤性黏膜中检测了 TERT 和其他端粒酶成分的表达，研究发现与相应的非肿瘤性粘膜相比，肿瘤性粘膜表现出 TERT 表达增加和端粒酶活性升高，在免疫组织化学中，在所有癌组织的肿瘤细胞核中均检测到 TERT 蛋白的强表达，而在非肿瘤性粘膜细胞以及基质成分(淋巴细胞除外)中，TERT 的表达弱或阴性。这些发现表明，与端粒酶活性相关的 TERT 表达增加可能是诊断胃癌的新标志。TFF2 是三叶因子家族(TFF)域肽中的成员之一，TFF 成员在肠粘膜防御和修复以及肿瘤发生中起作用，Leung 等人[17]的研究表明 TFF1 和 TFF2 表达之间存在显着相关性，可能与多步胃癌发生途径的早期阶段有关。TPX2 主要参与有丝分裂细胞周期，细胞凋亡，细胞周期以及激活蛋白激酶活性等生物过程，Heidebrecht 等人[18]研究发现 TPX2 (p100)可能被证明是一种更可靠的细胞增殖指标，并且与癌症的预后更加紧密相关。BLM 是一种蛋白质编码基因，参与 DNA 复制和修复。Broberg 等人[19]研究发现 BLM

复合物中的某些多态性是一般的癌症易感性标志物，并且可能参与了几种细胞类型的肿瘤转化。FOXM1基因编码的蛋白质是参与细胞增殖的转录激活因子，调节 DNA 复制和有丝分裂所必需的细胞周期基因的表达。在控制细胞增殖中发挥作用。Kim 等人[20]的研究结果表明在人类肿瘤衍生的众多细胞系中发现了Foxm1 水平升高。在多种人类肿瘤中还发现 Foxm1 蛋白的表达增加，包括肝细胞癌，基底细胞癌，乳腺腺癌，星形细胞瘤和胶质母细胞瘤，表明 Foxm1 调节各种人类癌症中的细胞增殖。IFNG 基因编码可溶性细胞因子，由被特定抗原或有丝分裂原激活的淋巴细胞产生。IFNG (干扰素 γ)除具有抗病毒活性外，还具有重要的免疫调节功能。它是巨噬细胞的有效活化剂，对转化细胞具有抗增殖作用，并且可以增强 I 型干扰素的抗病毒和抗肿瘤作用。MMP1 编码基质金属肽酶，Tsuchiya 等人[21]的研究表明 MMP-1 启动子中多态性位点的 2G 等位基因比 1G 等位基因具有更高的转录活性。携带 MMP-1 的 11q22 等位基因失衡经常在各种癌症中观察到，可能与晚期疾病有关，并且与癌症的侵袭和转移有关。CDK1 (细胞周期蛋白依赖性激酶 1)，通过调节中心体周期以及有丝分裂的发生，在控制真核细胞周期中起关键作用；通过与多个相间周期素相关联，促进 G2-M 过渡，并调节 G1 进程和 G1-S 过渡。Johnson 等人[22]的研究证实了 CDK2 和 CDK1 抑制剂在癌症治疗中的治疗潜力。CDK2 敲减导致 G1 积累，而 CDK1 耗竭导致 G2/M 减慢，而双 CDK1/2 耗竭导致抗雌激素敏感和耐药的细胞进一步 G2/M 积累和细胞死亡，从而确认 CDK2 和 CDK1 是乳腺癌的靶标癌症治疗。EXO1 (核酸外切酶 1)是参与错配修复系统的重要核酸酶，有助于维持基因组稳定性，调节 DNA 重组和介导细胞周期停滞。TIAM1 是与 T 淋巴瘤的转移相关的基因，该基因在细胞侵袭，转移和癌变中起重要作用。Liu 等人[23]通过研究抑制 TIAM1 表达对直肠癌细胞增殖和转移的影响，发现 TIAM1 在结直肠癌的转移中确实起着因果作用。Bau 等人[24]的研究发现 EXO1 中潜在的多态性可能会通过影响 EXO1 的修复活性而改变癌症风险，进而推测 EXO1 中的单核苷酸多态性(SNPs)可能与胃癌的风险有关。

4. 结论

胃癌是世界上最常见的癌症之一，尽管在过去一个世纪中，胃癌的发病率在全球范围内有所下降，但胃癌仍然是全球范围内的主要杀手。由于早期胃癌无特异性改变，大多患者发现已经处于进展期，手术、辅助放化疗的快速发展在一定程度上提高了胃癌患者的生存期，但由于化学药物抗性等原因胃癌患者的生存率比较低。因此识别与胃癌相关的分子标志物，为胃癌患者提供个体化治疗以提高治疗效果尤为重要。

DNA 拷贝数的变异也就意味着在 DNA 的复制过程中某些基因的增添或者删减，这必然会导致某些基因的表达发生改变，即高表达或者低表达。本文基于生物信息学的方法，对 375 例胃癌组织和 32 例正常组织的基因表达数据以及 DNA 拷贝数变异数据进行联合分析，筛选出 946 个共有的上调差异基因和 921 个共有的下调差异基因。对这些差异基因进行 GO 富集分析显示，81%的差异基因参与了 GO 术语的生物过程，包括趋化因子的信号通路，胞间信号的传导，细胞分化，DNA 复制，细胞增殖的正向调控以及细胞粘附等生物过程。已有相关研究证明了趋化因子的信号通路和细胞粘附等生物过程在胃癌的发生发展过程中的作用。KEGG 通路分析结果显示，这些基因显著富集于 PI3K-Akt 信号通路、ECM-受体相互作用、Wnt 信号通路以及胃酸分泌等。PI3K-Akt 信号通路在胃癌中的作用已有研究表明。通过 String 数据库构建差异基因的蛋白质相互作用网络，并结合生物信息软件 CytoScape 的插件 MCODE，找出了网络中联系最为紧密的两个子网络，联合 MalaCards 疾病数据库找出了 8 个与胃癌相关的基因，运用 CytoHubba 插件重建了与这八个基因联系最为紧密的网络。最后筛选出该网络中基因拷贝数与基因表达呈正相关的 26 个基因。

总之，本文为了更深入地了解与胃癌发生发展过程相关的基因和信号通路，整合分析了 TCGA 数据

库中胃癌的 RNA-Seq 和 DNA Copy Number 数据，运用生物信息学的方法筛选 26 个既发生差异表达又发生 DNA 拷贝数变异的驱动基因。GO 功能富集分析和 KEGG 通路分析显示，这些基因显著富集于 PI3K-Akt 信号通路、ECM-受体相互作用、Wnt 信号通路以及胃酸分泌等通路，这些基因和信号通路可能与胃癌的发生发展有关。其中的 19 个基因在胃癌发生发展过程中的作用已有相关文献报道且得到了临床证明，剩余的 AGT、PRKACG、F2、KNG1、GNGT1、SH3GL2 和 RHOV 等基因中，AGT、PRKACG、F2 和 KNG1 都参与了癌症途径，GNGT1 参与了 PI3K-Akt 信号通路，但在胃癌的发生发展过程中的作用尚无相关报道，有待进一步的研究和临床证明。

参考文献

[1] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A. and Jemal, A. (2018) Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA*: *A Cancer Journal for Clinicians*, **68**, 394-424. https://doi.org/10.3322/caac.21492

[2] Van Cutsem, E., Sagaert, X., Topal, B., Haustermans, K. and Prenen, H. (2016) Gastric Cancer. *The Lancet*, **388**, 2654-2664. https://doi.org/10.1016/S0140-6736(16)30354-3

[3] Tabernero, J., Hoff, P.M., Shen, L., Ohtsu, A., Shah, M.A., Cheng, K., Song, C.Y., Wu, H.Y., Eng-Wong, J., Kim, K., Kang, Y.-K. (2018) Pertuzumab plus Trastuzumab and Chemotherapy for HER2-Positive Metastatic Gastric or Gastro-Oesophageal Junction Cancer (JACOB): Final Analysis of a Double-Blind, Randomised, Placebo-Controlled Phase 3 Study. *The Lancet. Oncology*, **19**, 1372-1384. https://doi.org/10.1016/S1470-2045(18)30481-9

[4] The GASTRIC (Global Advanced/Adjuvant Stomach Tumor Research International Collaboration) Group (2010) Benefit of Adjuvant Chemotherapy for Resectable Gastric Cancer: A Meta-Analysis. *JAMA*, **303**, 1729-1737. https://doi.org/10.1001/jama.2010.534

[5] Luthra, R., Chen, H., Roy-Chowdhuri, S. and Singh, R.R. (2015) Next-Generation Sequencing in Clinical Molecular Diagnostics of Cancer: Advantages and Challenges. *Cancers* (*Basel*), **7**, 2023-2036.

[6] Robinson, M.D., Mccarthy, D.J. and Smyth, G.K. (2010) edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data. *Biogeoences*, **26**, 139-140. https://doi.org/10.1093/bioinformatics/btp616

[7] Huang, D.W., Sherman, B.T. and Lempicki, R.A. (2009) Bioinformatics Enrichment Tools: Paths toward the Comprehensive Functional Analysis of Large Gene Lists. *Nucleic Acids Research*, **37**, 1-13. https://doi.org/10.1093/nar/gkn923

[8] Huang, D.W., Sherman, B.T., Tan, Q., Kir, J., Liu, D., Bryant, D., Guo, Y.J., Stephens, R., Baseler, M.W., Clifford Lane, H. and Lempicki, R.A. (2007) DAVID Bioinformatics Resources: Expanded Annotation Database and Novel Algorithms to Better Extract Biology from Large Gene Lists. *Nucleic Acids Research*, **35**, W169-W175. https://doi.org/10.1093/nar/gkm415

[9] Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., Jensen, L.J. and Mering, C.V. (2019) STRING V11: Protein-Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-Wide Experimental Datasets. *Nucleic Acids Research*, **47**, D607-D613. https://doi.org/10.1093/nar/gky1131

[10] Dar, A.A., Belkhiri, A. and El-Rifai, W. (2009) The Aurora Kinase a Regulates GSK-3B in Gastric Cancer Cells. *Oncogene*, **28**, 866-875. https://doi.org/10.1038/onc.2008.434

[11] Zhou, J.-J., Chen, M.-L., Zhang, Q.-Z., Hu, J.-K. and Wang, W.-L. (2004) Coexpression of Cholecystokinin-B/Gastrin Receptor and Gastrin Gene in Human Gastric Tissues and Gastric Cancer Cell Line. *World Journal of Gastroenterology*, **10**, 791-794. https://doi.org/10.3748/wjg.v10.i6.791

[12] Yasumoto, K., Okamoto, S., Mukaida, N., Murakami, S., Mai, M. and Matsushima, K. (1992) Tumor Necrosis Factor Alpha and Interferon Gamma Synergistically Induce Interleukin 8 Production in a Human Gastric Cancer Cell Line through Acting Concurrently on AP-1 and NF-Kb-Like Binding Sites of the Interleukin 8 Gene. *The Journal of Biological Chemistry*, **267**, 22506-22511.

[13] Tanizaki, J., Okamoto, I., Takezawa, K., Tsukioka, S., Uchida, J., Kiniwa, M., Fukuoka, M. and Nakagawa, K. (2010) Synergistic Antitumor Effect of S-1 and HER2-Targeting Agents in Gastric Cancer with HER2 Amplification. *Molecular Cancer Therapeutics*, **9**, 1198-1207. https://doi.org/10.1158/1535-7163.MCT-10-0045

[14] Liang, Z.Y., Zeng, X., Gao, J., Wu, S.F., Wang, P., Shi, X.H., Zhang, J. and Liu, T.H. (2008) Analysis of EGFR, HER2, and TOP2A Gene Status and Chromosomal Polysomy in Gastric Adenocarcinoma from Chinese Patients. *BMC Cancer*, **8**, Article No. 363. https://doi.org/10.1186/1471-2407-8-363

[15] Chen, X.-H., Lan, B., Qu, Y., Zhang, X.-Q., Cai, Q., Liu, B.-Y. and Zhu, Z.-G. (2006) Inhibitory Effect of Polo-Like Kinase 1 Depletion on Mitosis and Apoptosis of Gastric Cancer Cells. *World Journal of Gastroenterology*, **12**, 29-35. https://doi.org/10.3748/wjg.v12.i1.29

[16] Yasui, W., Tahara, H., Tahara, E., Fujimoto, J., Nakayama, J.-I., Ishikawa, F., Ide, T. and Tahara, E. (1998) Expression of Telomerase Catalytic Component, Telomerase Reverse Transcriptase, in Human Gastric Carcinomas. *Cancer Science*, **89**, 1099-1103. https://doi.org/10.1111/j.1349-7006.1998.tb00502.x

[17] Leung, W.K., Yu, J., Chan, F.K.L., To, K.F., Chan, M.W.Y., Ebert, M.P.A., Ng, E.K.W., Sydney Chung, N.S.C., Peter, M. and Sung, J.J.Y. (2002) Expression of Trefoil Peptides (TFF1, TFF2, and TFF3) in Gastric Carcinomas, Intestinal Metaplasia, and Non-Neoplastic Gastric Tissues. *The Journal of Pathology*, **197**, 582-588. https://doi.org/10.1002/path.1147

[18] Heidebrecht, H.J., Buck, F., Steinmann, J., Sprenger, R., Wacker, H.H. and Parwaresch, R. (1997) P100: A Novel Proliferation-Associated Nuclear Protein Specifically Restricted to Cell Cycle Phases S, G2, and M. *Blood*, **90**, 226-233. https://doi.org/10.1182/blood.V90.1.226.226_226_233

[19] Broberg, K., Huynh, E., Schläwicke, E.K., Björk, J., Albin, M., Ingvar, C., Olsson, H. and Höglund, M. (2009) Association between Polymorphisms in *RMI*1, *TOP*3*A*, and *BLM* and Risk of Cancer, a Case-Control Study. *BMC Cancer*, **9**, Article No. 140.https://doi.org/10.1186/1471-2407-9-140

[20] Kim, I.-M., Ackerson, T., Ramakrishna, S., Tretiakova, M., Wang, I.-C., Kalin, T.V., Major, M.L., Gusarova Galina A., Yoder, H.M., Costa, R.H. and Kalinichenko, V.V. (2006) The Forkhead Box M1 Transcription Factor Stimulates the Proliferation of Tumor Cells during Development of Lung Cancer. *Cancer Research*, **66**, 2153-2161. https://doi.org/10.1158/0008-5472.CAN-05-3003

[21] Tsuchiya, N., *et al*. (2009) Clinical Significance of a Single Nucleotide Polymorphism and Allelic Imbalance of Matrix Metalloproteinase-1 Promoter Region in Prostate Cancer. *Oncology Reports*, **22**, 493-499. https://doi.org/10.3892/or_00000462

[22] Johnson, N., Bentley, J., Wang, L.Z., Newell, D.R., Robson, C.N., Shapiro, G.I. and Curtin, N.J. (2010) Pre-Clinical Evaluation of Cyclin-Dependent Kinase 2 and 1 Inhibition in Anti-Estrogen-Sensitive and Resistant Breast Cancer Cells. *British Journal of Cancer*, **102**, 342-350. https://doi.org/10.1038/sj.bjc.6605479

[23] Liu, L., Zhang, Q.L., Zhang, Y.F., Wang, S. and Ding, Y.Q. (2006) Lentivirus-Mediated Silencing of *Tiam*1 Gene Influences Multiple Functions of a Human Colorectal Cancer Cell Line. *Neoplasia*, **8**, 917-924. https://doi.org/10.1593/neo.06364

[24] Bau, D.T., Wang, H.C., Liu, C.S., Chang, C.L., Chiang, S.Y., Wang, R.F., Tsai, C.W., Lo, Y.L., Hsiung, C.A., Lin, C.C. and Huang, C.Y. (2009) Single-Nucleotide Polymorphism of the Exo1 Gene: Association with Gastric Cancer Susceptibility and Interaction with Smoking in Taiwan. *Chinese Journal of Physiology*, **52**, 411-418.