doi:10. 3969/j.issn.1672-5565.2015.03.05

癌症 DNA 甲基化调控位点的识别

韦云真[1],刘晓娟[2],王 芳[1],苏建忠[1],张岩[1*],刘洪波[1*]
(1. 哈尔滨医科大学生物信息科学与技术学院,哈尔滨 150081;
2.哈尔滨医科大学附属第一医院康复医学科,哈尔滨 150001)

摘 要:DNA 甲基化是一种重要的表观遗传学修饰,在基因的转录调控方面具有重要的作用。异常的 DNA 甲基化可以导致癌症等复杂疾病发生,癌基因相关的 DNA 甲基化调控位点的识别对于解析癌症的发生发展机制及识别新的癌症标记具有重要意义。本研究通过整合 The Cancer Genome Atlas(TCGA)的泛癌症基因组的高通量甲基化谱和基因表达谱,识别癌基因相关的 DNA 甲基化调控位点。对于每种癌症分批次计算 CpG 位点甲基化与相关基因表达之间的相关性,并筛选调控下游基因的 CpG 位点(包括强调控位点、弱调控位点和不调控位点),结果表明仅有一半的 CpG 位点对下游基因具有调控作用;对癌症间共享的调控位点的分析发现不同癌症间共享的调控位点不尽相同,表明癌症特异的甲基化调控位点的存在。进一步地,对差异甲基化和差异表达基因的功能富集分析揭示了受甲基化调控的基因确实参与了癌症发生发展相关的功能。本研究的结果是对当前甲基化调控位点集的重要补充,也是识别癌症新型分子标记特征的重要资源。

关键词:DNA 甲基化;基因表达;转录调控;癌症
中图分类号:R73;Q7 **文献标志码**:B **文章编号**:1672-5565(2015)03-170-09

Identification of cancer DNA methylation regulatory sites

WEI Yunzhen[1], LIU Xiaojuan[2], WANG Fang[1], SU Jianzhong[1], ZHANG Yan[1*], LIU Hongbo[1*]
(1. *College of Bioinformatics Science and Technology*, *Harbin Medical University*, *Harbin* 150081, *China*;
2. *Department of Rehabilitation*, *The First Affiliated Hospital of Harbin Medical University*, *Harbin* 150001, *China*)

Abstract: DNA methylation is an important epigenetic modification, which plays an important role in the regulation of gene transcription. Abnormal DNA methylation may lead to cancer and disease, and identifying oncogene-related DNA methylation gene regulatory sites is important for the development of mechanisms to resolve the occurrence of cancer and identify new cancer markers. In this study, we integrate high-throughput DNA methylation profiling and gene expression profiling of pan-cancer genome in TCGA, then identify oncogene-related DNA methylation regulation sites. For each cancer, we calculate the correlation between methylation of CpG sites and gene expression, and filter the CpG sites, which regulate downstream genes (including strong regulatory sites, weak regulatory sites and not regulatory sites). The results show thatonly half of the CpG sites regulate the downstream genes. Analyzing of regulatory sites that is shared between cancers show that regulatory sites are not necessarily the same in different cancer, and the presence of cancer-specific methylation regulatory sites. Moreover, gene function enrichment analysis of differential DNA methylation and differentially expressed genes show that genes regulated by methylation are indeed involved in the development of cancer-related functions. The results of this study are an important supplementation to the current DNA methylation regulatory sites set, and an important resource to identify new molecular markers characteristics of cancer.

Keywords: DNA methylation; Gene expression; Transcriptional regulation; Cancer

收稿日期:2015-05-27;**修回日期**:2015-07-20.
基金项目:国家自然科学基金项目(61403112, 31371334)。
作者简介:韦云真,女,本科生,研究方向:计算表观遗传学;E-mail: weiyunzhen@ yeah.net.
*通信作者:张岩,女,教授,研究方向:计算表观遗传学、生物信息学;E-mail:tyozhang@ ems.hrbmu.edu.cn;
刘洪波,男,讲师,研究方向:计算表观遗传学、生物信息学;E-mail:hongbo919@ gmail.com.

DNA 甲基化是一种重要的表观遗传学修饰，在 CpG 岛（DNA 的 CG 序列密集区）上发生，对调控转录基因具有重要的作用[1-3]。甲基化位点可随 DNA 的复制而遗传，因为 DNA 复制后，甲基化酶可将新合成的未甲基化的位点进行甲基化[4]。DNA 的甲基化可引起基因的失活。如 CpG 岛位于某基因的启动子区域，CpG 岛的甲基化会显著降低甚至完全沉默该基因的转录，继而影响蛋白的表达。CpG 岛的甲基化程度越高，基因表达的程度越低。目前对于甲基化调控基因表达，以及进一步的生物学影响的研究很多[5]。然而对具体的基因，尚没有一个完整的甲基化调控区域的图谱。

最近几年来，表观遗传学领域发展十分迅速。DNA 甲基化修饰就是一个非常重要的部分，参与基因表达调控、转座子沉默、X 染色体失活、基因印记、以及癌症发生等重要生物学过程[6-8]。近年来随着研究技术和方法的进步，全基因组 DNA 甲基化的研究广泛兴起，很多物种的全基因组甲基化图谱破译了出来，DNA 甲基化全局水平的研究不仅有利于宏观层面上了解 DNA 甲基化的规律和特性，同时也为深入分析 DNA 甲基化生物学调控及功能奠定了基础。如今，在当前领域已经取得了一些进展，例如发现了 DNA 甲基化酶的 DNMT 家族，并且对其作用机制和生理功能进行了一些研究[9-11]。甲基化与癌症的发生有关系[12-14]。研究发现抑癌基因启动子高度甲基化后，可以令这些基因表达受到抑制，同癌症的发生有着十分密切的关系，从而研究 DNA 甲基化抑制剂的使用，将有助于预防人类肿瘤的发生。

基于高通量的 DNA 甲基化数据[15]，研究 DNA 甲基化与基因表达之间的关系，并建立生物学测度筛选对基因表达之间的关系，筛选对基因具有调控作用的 DNA 甲基化区域，最后绘制基因组范围内参与基因调控的 DNA 甲基化区域。研究成果将有助于对表观遗传调控机制更深入理解。

1　材料与方法

1.1　材料

研究使用 The Cancer Genome Atlas（TCGA）上同时具有甲基化数据和表达数据的所有 27 K 高通量癌症数据[16]。如表 1 所示，根据以上挑选条件，总计 11 癌症符合条件，15 套数据，81 个处理批次（Batch）。不同癌症内包含着不同的样本数与处理批次。其中 BRCA，OV 与 READ 癌症数据同时包含有癌症样本与正常样本，如表 1 所示。基因表达数据与甲基化数据样本数数量一致，并且是一一对应的，基因表达部分使用 level2 数据，level2 数据内容为探针名-取 log2 后的表达值。甲基化部分使用的是 level3 数据，level3 数据内容为甲基化位点名-甲基化值。

表 1　研究采用的数据

Table1　Data of the research

名称	批次	样本	癌症	正常
BRCA	4	334	313	21
GBM	18	577	577	0
KIRC	2	71	71	0
KIRP	1	16	16	0
LAML	3	191	191	0
LGG	1	27	27	0
LUAD	2	30	30	0
LUSC	10	267	267	0
OV	31	1 190	1 174	16
READ	7	71	68	3
UCEC	2	54	54	0

1.2　方　法

1.2.1　数据预处理

将 Methylation、Expression 分批次进行数据的标准化。对于每一个甲基化位点，我们按不同的 Batch 计算这个甲基化位点一一对应样本的 Methylation 部分及 Expression 部分数据的 P 值及皮尔森相关系数（PCC），进行 DNA 甲基化-基因调控关系定量。合并各 Batch 的结果并进行进一步的分析。

1.2.2　癌症甲基化与表达值的 PCC 分布曲线绘制

保留皮尔森相关系数显著（$P<0.05$）的 CpG 位点-基因对，并利用 R 语言里的 ggplot2 包进行分布曲线的绘制。

为了验证在 P 值取不同阈值的情况下，PCC 值的分布是否是稳定的，取不同的显著性 P 值下进行重复研究，显著性阈值分别取 $P<1.0\times10^{-2}$，$P<1.0\times10^{-3}$，$P<1.0\times10^{-4}$，$P<1.0\times10^{-5}$，$P<1.0\times10^{-6}$，$P<1.0\times10^{-7}$。并分别绘制 PCC 值分布曲线图。

1.2.3　Batch 间相关性分析

使用预处理数据研究癌症 Batch 间的相关性，我们进行了 Batch 间的相关性分析，对于每一个 Batch，我们认为 P 小于 0.05 时的 PCC 值是有用的，进行保留，而 P 值大于等于 0.05 时的 PCC 值认为是没有用的，将这时的 PCC 值更改为 0。接着，将所有的 P 值列删除，只保存 PCC 值。计算两两 Batch 间的皮尔森相关系数。用 cluster 进行双向聚类。并将聚类结果用 TreeView 进行可视化。

1.2.4　甲基化位点与 Batch 相关性分析

对于每一个 Batch，当 P 值小于 0.05 时，PCC 值有效，保留原值，当 P 值大于等于 0.05 时，认为是不

显著的,PCC 值改为 0。这样,做成一张横向为 81 个癌症 Batch 名,纵向为 25 851 个甲基化 cg 位点的表格。通过这个表格,可以看出 cg 位点与每一个癌症 Batch 的关系。使用 cluster 软件,对这个表格进行筛选,筛选出来的甲基化位点数为 2 420 个。将这个横向为 81 个 Batch,纵向为 2 420 个甲基化位点的表格用 cluster 进行欧式距离的双向聚类。

1.2.5 相关性数据离散化分析

对甲基化位点与 Batch 相关性分析做进一步分析,把 PCC 值分为五个区间,这五个区间分别是-1~-0.4,-0.4~-0.1,-0.1~-0.1,0.1~0.4,0.4~1.0。其中,-1~-0.4 区间代表强负相关;-0.4~-0.1区间代表弱负相关;-0.1~0.1 区间代表不相关;0.1~0.4 区间代表弱正相关;0.4~1.0 的区间,代表强正相关。按照这五个部分所占百分比做成饼图。

1.2.6 筛选强、弱以及无相关位点-转录本对

筛选各癌症强相关位点数与弱相关位点数。对于每一个位点-转录本对,样本总数为 N 个,大于等于 0.4 的样本值有 X 个,如果 X/N 大于等于 0.5,则认为该位点对癌症是有强调控作用的;如果大于 0.1 的样本值有 Y 个,如果 Y/N 大于等于 0.5,则认为该位点对癌症是有弱调控作用的。

1.2.7 绘制综合癌症数据与所有单个数据的韦恩图

将挑选出来的综合癌症数据,以及 BRCA 癌症数据,GBM 癌症数据,KIRC 癌症数据,KIRP 癌症数据的基因转录本,画韦恩图。

1.2.8 GO 注释

将筛选出来的强相关位点,弱相关位点对应的癌症关联的转录本,放入 DAVID 里进行 GO 注释,查看其生物学途径,分子功能,细胞组件。

表达数据差异筛选与甲基化数据差异筛选并进行 GO 注释。在下载下来的 11 套数据中,其中 BRCA 癌症数据,OV 癌症数据,READ 癌症数据中同时含有正常样本与癌症样本。对这三套数据,进行癌症样本表达数据和正常样本表达数据的差异筛选,使用 SAM 方法进行差异筛选。再分别做这三套癌症的甲基化数据的差异基因筛选。把筛选出来的差异表达数据转录本与差异甲基化数据转录本放入 DAVID 里进行 GO 注释,查看与其相关的生物学途径,分子功能,细胞组件,并进行 GO 分类富集分析。在做各癌症 GO 注释描述及 GO 分类富集分析的时候,认为 P-Value 及 Benjiaminj 值小于 0.01 时是显著的。

2 结果与讨论

2.1 癌症甲基化与表达值的 PCC 分布

基于 TCGA 的高通量的泛癌 DNA 甲基化和基因表达谱数据,我们利用皮尔森相关系数(PCC)对 CpG 位点对基因表达调控作用进行了定量。如图 1 所示,每一条曲线代表的是一个癌症 Batch 的 PCC 分布情况,图中共有 81 条重叠曲线。在弱负相关与弱正相关处出现两个峰值。当 PCC 呈现弱负相关时,基因出现表达,这符合我们所说的,甲基化程度低,表达程度高,然而当 PCC 值呈现正负相关时,也出现了一个峰值,但右边的峰值略低于左边。另外,图 1 出现了与其他 Batch 不相似的 Batch 曲线,粉色*的曲线为 READ_7_Bacth_1758,紫色*的曲线为 OV_7_Bacth_1141数据,蓝色*的曲线为 133_OV_1138 数据。

根据选取六个不同的 *P* 值值域,画出了六个 PCC 分布曲线图,如图 2 所示,在 *P* 值取不同临界值的情况下,绝大多数癌症 Batch 的 PCC 分布曲线没有发生明显的变化,个别的曲线随着 *P* 值的变化而发生改变,这证明了 PCC 的分布情况是比较稳定的。

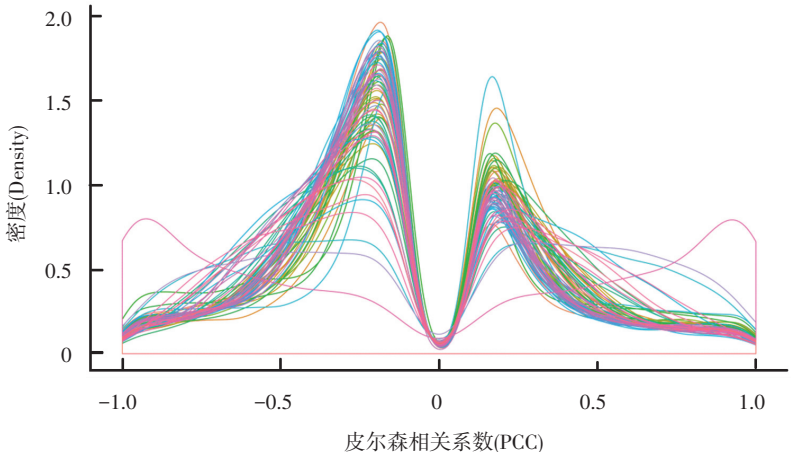

图 1 当 *P*=0.05 时癌症甲基化与表达值的 PCC 分布

Fig.1 When *P*=0.05, cancer methylation and expression of value distribution of the PCC

*:图中颜色标注见电子版(http://swxxx.alljournals.cn/index.aspx)(2015 年第 3 期)。

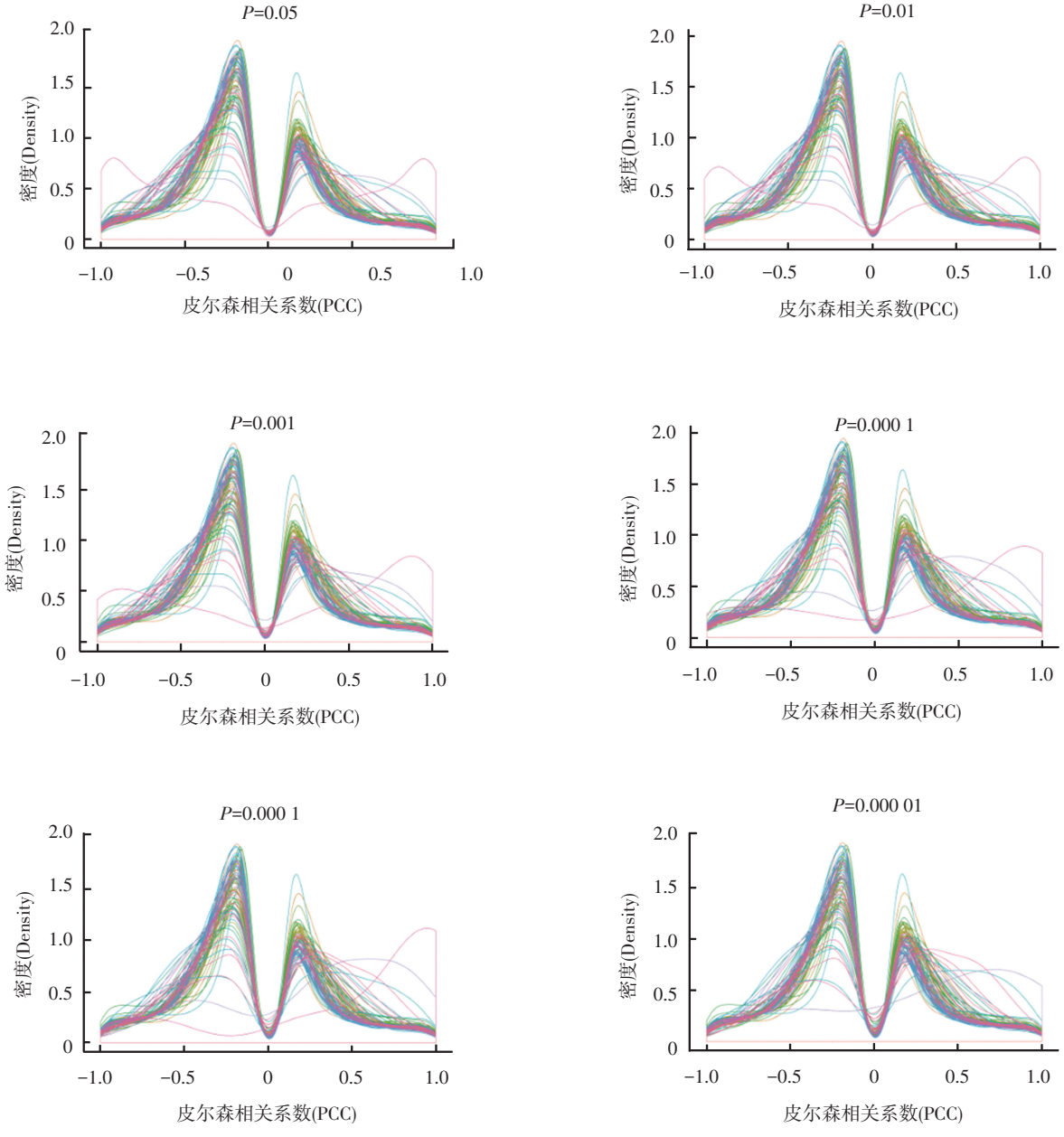

图 2　取不同 *P* 值情况下 PCC 的分布

Fig.2　When *P* value take different cases, the distribution of the PCC

*：图中颜色标注见电子版(http://swxxx.alljournals.cn/index.aspx)(2015 年第 3 期)。

2.2　Batch 间相关性分析

红色*越深，代表着相似性越显著。红色*最显著的斜对角线是每个 Batch 和自身的相似性，因此最为显著。从 Batch 间相关性分析的聚类可以看到，图 3 可视图呈现出块状聚集的分布，处于相同癌症中的 Batch 的聚类效果比较显著。而对于不同癌症间的 Batch，聚类效果不明显。不同癌症之间没有明显的联系。

2.3　甲基化位点与 Batch 相关性分析

使用 cluster 进行欧式距离的双向聚类，再用 treeview 进行可视化，得到图 4。红色*的部分代表着某个位点的甲基化对该癌症 Batch 有调控的作用。红色*越深表示调控的作用越显著。绿色*的部分代表着这个位点的甲基化对该 Batch 的调控不显著。从横向来看，分析的是这 2 420 个 CpG 位点调控着哪些癌症 Batch，从纵向来看，分析的是 Batch 共享哪些 CpG 位点的调控。由图 4 可以看到，一些 CpG 位点显著调控着所有的 Batch，为所有癌症所共享；一些 CpG 位点显著着调控个别 Batch，而对其他的 Batch 调控是不显著的，是癌症特异的；有些 CpG 位点在图上显示对所有的 Batch 都没有显著的调控，这是 27K 数据一个不足的地方，位点信息仍存在着缺失。这个图绘制了甲基化区域的调控图谱。

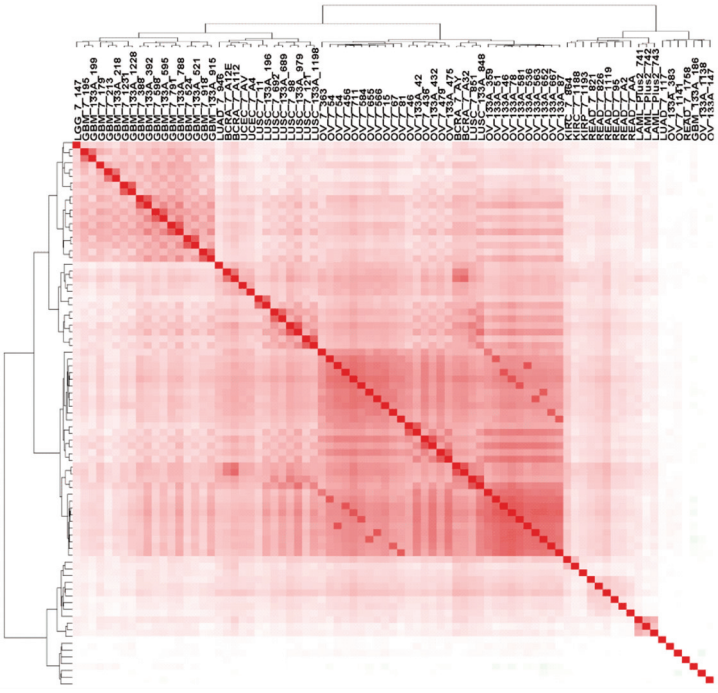

图 3 Batch 间相关性分析

Fig.3 Correlation analysis of Batch

*:图中颜色标注见电子版(http://swxxx.alljournals.cn/index.aspx)(2015 年第 3 期)。

2.4 相关性数据离散化分析

根据强负相关,弱负相关,无相关,弱正相关,强正相关这五个部分个数,画出图 5 这个饼图。浅蓝* 部分为强负相关区域,橙色* 部分为弱负相关区域,灰色* 部分为无相关区域,黄色* 部分为弱正相关区域,深蓝* 部分强正相关区域。从饼图上可以很明确的看出每一个部分所占的百分比。从饼图上可以看到大多数 PCC 值都是无相关的。弱相关部分远多于强相关部分。

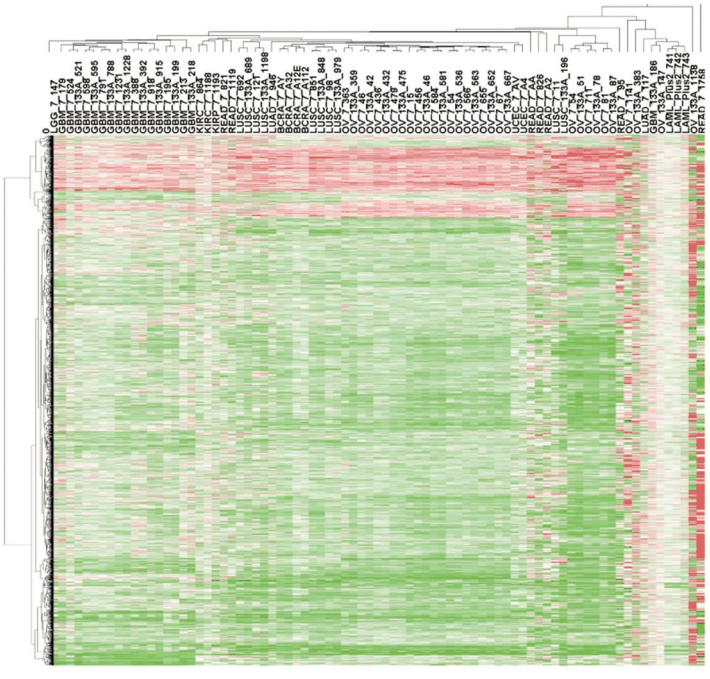

图 4 甲基化位点与 Batch 相关性分析

Fig.4 Correlation analysis of methylation site and Batch

注:横轴为筛选出来的 2 420 个甲基化位点,纵轴为 81 个癌症 Batch。

Notes: X axis for 2 420 methylation site, Y axis for Batch 81 cancer.

*:图中颜色标注见电子版(http://swxxx.alljournals.cn/index.aspx)(2015 年第 3 期)。

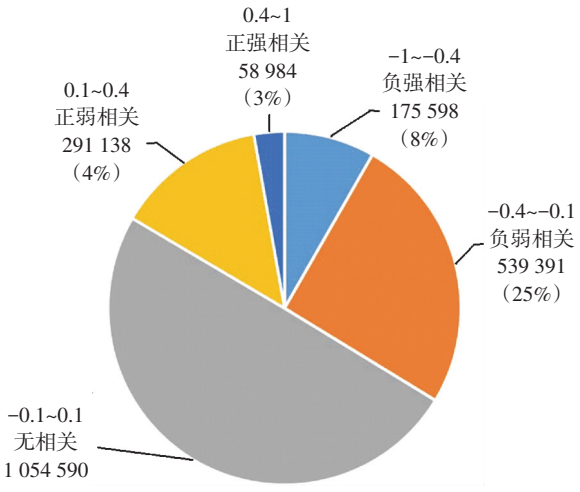

图 5　不同区间 PCC 值范围所占百分比

Fig.5　Different interval PCC percentage value range

2.5　筛选强相关、弱相关以及无相关位点-转录本对

对综合了所有癌症数据的 excel 表进行筛选之后，如图 6 所示，在综合了所有癌症批次数据下挑选出来的强相关位点有 186 个、弱相关位点 16 280 个，与无相关位点 25 280 个。其中无相关位点占 61%，是绝大多数，弱相关位点占 39%，而强相关位点只有 186 个，只有一小部分。对于单个的癌症而言，无相关位点同样占据了绝大多数，弱相关位点多于强相关位点。

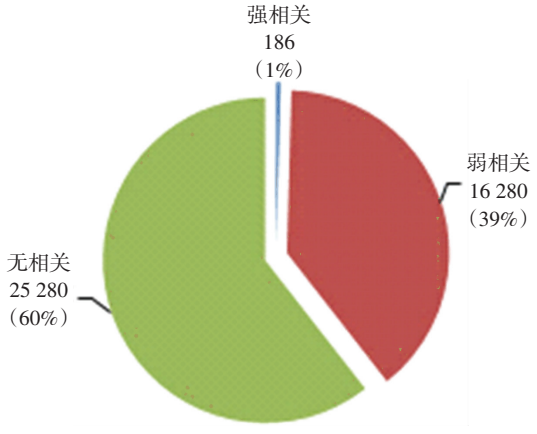

图 6　强相关位点、弱相关位点、无相关位点所占百分比

Fig.6　Related sites, weak related sites, no relevant sites for percentage

2.6　绘制综合癌症数据与单个癌症数据的韦恩图

进一步我们研究了各癌症间共享的强相关位点的数量（见图 7），可见，有些位点和其他癌症都有关联，并不是局限于某个癌症，为这几个癌症共享；有些转录本被若干个癌症所共享；而有些转录本是癌症特异的，只与该癌症相关，不调控其他的癌症。例如基因 Ddx43 的转录本 NM_018665，被这所有的五个数据集合共享，Ddx43 与个体死亡有关。基因 Dynlrb2 的转录本 NM_130897，同时被 BRCA 与 KIRP 共享，Dynlrb2 调控动力蛋白。基因 LAPTM5 的转录本 NM_006762，同时被 GBM 与 KIRC 共享，LAPTM5 和溶酶体 multispanning 膜蛋白 5 有关。基因 kazald1 的转录本 NM_030929 同时被 BRCA，GBM，KIRC 共享，kazald1 和 Kazal-type 丝氨酸肽酶抑制结构域 1 有关。基因 SLC7A2 的转录本 NM_001008539，为 BRCA 所特有，SLC7A2 与溶质载体家族 7 有关。fgf1 基因的转录本 NM_033136，为 GBM 所特有，fgf1 与纤维原细胞生长因子 1 有关。基因 EPHA7 的转录本 NM_004440，为 KIRC 特有，EPHA7 与 EPH 受体 7 有关。基因 SERPINE2 的转录本 NM_006216，为 KIRP 特有，SERPINE2 与 serpin 肽酶抑制剂有关。分别研究调控所有数据的癌症基因，以及癌症特异的基因。

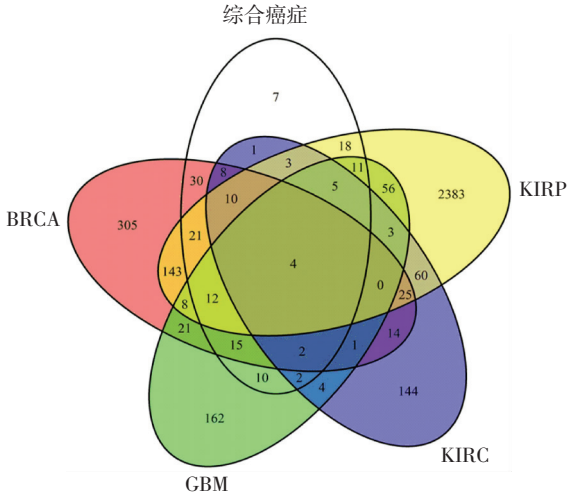

图 7　综合癌症数据、BRCA、GBM、KIRC、KIRP 数据韦恩图

Fig.7　Comprehensive cancer data, BRCA, GBM, KIRC, KIRP for venn

2.7　筛选强、弱以及无相关位点-转录本对

筛选出所有单个癌症数据及所有综合数据的符合条件的弱相关位点对、强相关位点对以及无相关位点对。在所有癌症中强相关位点数有 186 个，弱相关位点数 16 280 个。

表 2　筛选出的各癌症强相关位点数与弱相关位点数

Table 2　Select all the cancer related points and weak related points

癌症名称	强相关位点数	弱相关位点数
所有癌症	186	16 280
BRCA	746	9 838
GBM	409	13 501
KIRC	307	7 838
KIRP	3 301	13 261
LAML	348	12 619
LGG	2 311	13 233
LUAD	666	8 932
LUSC	928	12 881
OV	678	14 523
READ	2 644	15 739
UCEC	628	8 321

2.8 筛选强相关位点、弱相关位点的甲基化与表达数据的差异位点数

筛选强相关位点、弱相关位点的甲基化与表达数据的差异位点数,在表达数据里差异表达的位点数,在甲基化数据里也是差异的。

表 3 强相关位点、弱相关位点的甲基化差异位点数与表达数据差异位点数

Table 3 Methylation and expressing differences sites of strong related sites and weak related sites

数据名称	甲基化差异位点数	表达数据差异位点数
BRCA 强相关	139	77
OV 强相关	140	54
READ 强相关	33	97
BRCA 弱相关	3 840	4 731
OV 弱相关	974	1 000
READ 弱相关	71	3 583

2.9 各癌症 GO 注释描述及 GO 分类富集分析

对有癌症样本与正常样本的三套癌症数据进行表达数据的差异筛选。其中乳腺癌强相关表达数据差异位点数为 77 个,弱相关表达数据差异位点数为 4 731 个;卵巢癌强相关表达数据差异位点数为 54 个,弱相关表达数据差异位点数为 1 000 个;直肠癌强相关表达数据差异位点数为 97 个,弱相关表达数据差异位点数为 3 583 个。从基因层面分析各个癌症的转录本,将癌症转录本以癌症为单位放入 DAVID 中进行 GO 注释描述、GO 分类富集分析。

如表 4 所示,BRCA 癌症里有 4 808 个转录本,READ 癌症里有 3 680 个转录本,OV 癌症有 1 026 个转录本。对乳腺癌癌症在生物学过程 BP_1 层面的 GO 注释,发现乳腺癌表达数据转录本富集在细胞过程等基本生物学过程上(P,Benjamini<0.01),除此之外,可以看到这些基因转录本对乳腺癌癌症有发育的作用,影响癌细胞的增殖;使得癌细胞有附着力,能附着在组织或者器官上;影响着再增殖的过程,使得癌细胞无限繁殖继而继续生长;富集的基因有能使癌细胞移动的能力,使得癌细胞扩散到其他组织中,并且会导致死亡等。

如表 5 所示,对于卵巢癌癌症表达数据在生物学过程 BP_2 层面的 GO 注释(P,Benjamini<0.01),发现卵巢癌癌症表达数据集合不仅分别富集在细胞周期,细胞分裂,细胞凋亡等功能上,而且有些基因对生物学过程,细胞过程有负调控作用。一些基因注释在细胞扩散的功能上,解释了卵巢癌癌症癌症细胞在病人身上发生扩散和转移的现象。在表中,还可以看到有些基因注释为刺激细胞产生反应,也就是说当癌症发生时,这些基因的作用为刺激癌细胞,使得癌细胞产生各种机体反应。

表 4 BRCA 癌症在 BP_1 层面的 GO 注释描述

Table 4 GO annotation description of BRCA in BP_1

层次类别	功能类别	*P* 值	本杰明校正值
GOTERM_BP_1	cellular process	5.1×10^{-30}	1.1×10^{-28}
GOTERM_BP_1	cellular component organization	2.7×10^{-23}	3.0×10^{-22}
GOTERM_BP_1	developmental process	6.9×10^{-19}	5.0×10^{-18}
GOTERM_BP_1	cellular component biogenesis	1.2×10^{-12}	6.3×10^{-12}
GOTERM_BP_1	death	9.4×10^{-12}	4.1×10^{-11}
GOTERM_BP_1	biological adhesion	2.1×10^{-7}	7.7×10^{-7}
GOTERM_BP_1	locomotion	3.7×10^{-7}	1.2×10^{-6}
GOTERM_BP_1	metabolic process	4.0×10^{-7}	1.1×10^{-6}
GOTERM_BP_1	localization	2.3×10^{-5}	5.7×10^{-5}
GOTERM_BP_1	growth	2.5×10^{-5}	5.4×10^{-5}
GOTERM_BP_1	multi-organism process	7.5×10^{-4}	1.5×10^{-3}
GOTERM_BP_1	reproduction	1.1×10^{-3}	2.1×10^{-3}
GOTERM_BP_1	reproductive process	1.3×10^{-3}	2.2×10^{-3}
GOTERM_BP_1	multicellular organismal process	1.5×10^{-3}	2.4×10^{-3}
GOTERM_BP_1	establishment of localization	2.2×10^{-3}	3.3×10^{-3}

表 5　OV 癌症在 BP_2 层面的 GO 注释描述

Table5　GO annotation description of OV in BP_2

层次类别	功能类别	*P* 值	本杰明校正值
GOTERM_BP_2	cell cycle process	7.4×10^{-23}	1.4×10^{-20}
GOTERM_BP_2	cell cycle	8.0×10^{-23}	7.8×10^{-21}
GOTERM_BP_2	organelle organization	9.5×10^{-19}	6.1×10^{-17}
GOTERM_BP_2	cell division	3.7×10^{-16}	1.6×10^{-14}
GOTERM_BP_2	microtubule-based process	7.4×10^{-9}	2.9×10^{-7}
GOTERM_BP_2	chromosome segregation	1.9×10^{-8}	6.2×10^{-7}
GOTERM_BP_2	anatomical structure development	6.1×10^{-7}	1.7×10^{-5}
GOTERM_BP_2	cell proliferation	3.3×10^{-6}	8.0×10^{-5}
GOTERM_BP_2	cellular response to stimulus	1.2×10^{-5}	2.5×10^{-4}
GOTERM_BP_2	negative regulation of cellular process	1.5×10^{-5}	3.0×10^{-4}
GOTERM_BP_2	negative regulation of biological process	1.6×10^{-5}	2.8×10^{-4}
GOTERM_BP_2	multicellular organismal development	5.1×10^{-5}	8.3×10^{-4}
GOTERM_BP_2	anatomical structure morphogenesis	5.3×10^{-5}	8.0×10^{-4}
GOTERM_BP_2	cell death	6.3×10^{-5}	8.8×10^{-4}
GOTERM_BP_2	interspecies interaction between organisms	2.3×10^{-4}	3.0×10^{-3}
GOTERM_BP_2	cellular component assembly	3.0×10^{-4}	3.6×10^{-3}
GOTERM_BP_2	DNA packaging	3.8×10^{-4}	4.4×10^{-3}

对于直肠癌癌症表达癌症在生物学过程 BP_1 层面的 GO 注释(P,Benjamini<0.01),可以看到直肠癌表达数据基因集合注释在生长,增值的功能上,在富集基因的作用下,促进癌细胞不断增值,发育。注释在粘附的功能上,使得癌细胞粘附在器官或组织上,得以进一步的分裂,增值,又可以看到,癌细胞增长的同时,机体对刺激发生了反应,又促进了免疫学的过程。注释在运动的功能上,这些富集基因的功能促进了癌细胞的转移和扩散到其他器官和组织上。

表 6　READ 癌症在 BP_1 层面的 GO 注释描述

Table 6　GO annotation description of READ in BP_1

层次类别	功能类别	*P* 值	本杰明校正值
GOTERM_BP_1	developmental process	4.7×10^{-33}	1.0×10^{-31}
GOTERM_BP_1	biological adhesion	3.3×10^{-17}	3.7×10^{-16}
GOTERM_BP_1	multicellular organismal process	4.8×10^{-16}	3.2×10^{-15}
GOTERM_BP_1	cellular component organization	1.2×10^{-11}	6.8×10^{-11}
GOTERM_BP_1	cellular process	4.8×10^{-11}	2.1×10^{-10}
GOTERM_BP_1	death	2.2×10^{-9}	8.2×10^{-9}
GOTERM_BP_1	biological regulation	2.9×10^{-9}	9.1×10^{-9}
GOTERM_BP_1	immune system process	6.5×10^{-9}	1.8×10^{-8}
GOTERM_BP_1	locomotion	7.5×10^{-7}	1.8×10^{-6}
GOTERM_BP_1	response to stimulus	1.8×10^{-5}	3.9×10^{-5}
GOTERM_BP_1	growth	1.6×10^{-4}	3.1×10^{-4}
GOTERM_BP_1	localization	1.8×10^{-4}	3.3×10^{-4}
GOTERM_BP_1	cellular component biogenesis	8.9×10^{-4}	1.5×10^{-3}
GOTERM_BP_1	rhythmic process	4.0×10^{-3}	6.2×10^{-3}

3 结论与讨论

3.1 结论

本研究通过整合 TCGA 的泛癌症基因组的高通量甲基化谱和基因表达谱,识别癌基因相关的 DNA 甲基化调控位点;结果表明仅有一半的 CpG 位点对下游基因具有调控作用;且存在癌症特异的甲基化调控位点;并揭示这些位点调控的基因确实参与了癌症发生发展相关的功能。

3.2 讨论

不同癌症 Batch 的 PCC 值分布曲线是相似的并且稳定,但是在直肠癌与卵巢癌里有 3 套 Batch 的 PCC 分布曲线出现异常,这可能是数据量过少或者数据不完善的原因造成的。

癌症的强相关位点数量远小于弱相关位点,并且大部分的位点是无相关的,这代表在 27 K 芯片测的启动子区域数据是有许多遗漏的,仍需完善。研究分析可知,有一些位点稳定的调控着所有的癌症,与所有的癌症都有这关联。有些位点是癌症特异的,只与这些癌症有关联,可以进一步分析这些启动子区域的位点是如何影响这些癌症的发生。有些位点与若干个癌症相关联,有的位点没看出对其他癌症有调控作用。

被所有癌症共享的基因很少,而癌症的发生往往不是只受到一个基因的影响,而是分别由几个基因共同作用而产生的。同时癌症也也受到特异的基因的影响。后续研究可以分别挑选这些不同类别基因进行研究。

通过将筛选出的癌症差异基因,放入 DAVID 中,做以癌症为单位的三套癌症表达数据的 GO 注释,得出结论,这些筛选出来的差异基因在生物学过程上,确实是与各癌症有着密切的关联。本研究的结果是对当前甲基化调控位点集的重要补充,也是识别癌症新型分子标记特征的重要资源。

参考文献(References)

[1] JONES P A. Functions of DNA methylation: islands, start sites, gene bodies and beyond[J]. Nature Reviews Genetics, 2012, 13(7):484-492.

[2] FAN G. DNA methylation and its basic function[J]. Neuropsychopharmacology Reviews, 2012,38(1): 23-38.

[3] SHAMES D S. DNA methylation in health, disease, and cancer[J]. Current Molecular Medicine, 2007,7(1):85-102(18).

[4] DAY J J, SWEATT J D. DNA methylation and memory formation[J]. Nature Neuroscience, 2010, 13(11): 1319-1323.

[5] WU H, ZHANG Y. Reversing DNA methylation: mechanisms, genomics, and biological functions[J]. Cell, 2014, 156:45-68.

[6] REA M,ZHENG W, CHEN M, et al. Histone H1 affects gene imprinting and DNA methylation in Arabidopsis[J]. Plant Journal, 2012, 71(5): 776-786.

[7] ZALA D, HINCKELMANN M V, YU H, et al. Vesicular glycolysis provides on-board energy for fast axonal transport[J]. Cell, 2013, 152(3):479-491.

[8] SUN H S, KENNEDY P J, NESTLER E J. Epigenetics of the depressed brain: role of histone acetylation and methylation[J]. Neuropsychopharmacology Official Publication of the American College of Neuropsychopharmacology, 2013,38(1):124-137.

[9] LRY T J, LI D, WALTER M J, et al. DNMT3A mutations in acute myeloid leukemia[J]. New England Journal of Medicine, 2010, 363(25): 2424-2433.

[10] RUSICIO A D, EBRALIDZE A K, BENOUKRAF T, et al. DNMT1-interacting RNAs block gene-specific DNA methylation[J]. Nature, 2013, 503(7476):371-376.

[11] GUO X, WANG L, LI J, et al. Structural insight into autoinhibition and histone H3-induced activation of DNMT3A [J]. Nature, 2015, 517(7536):640-644.

[12] COPPIETERS N, DIERIKS B V, LILL C, et al. Global changes in DNA methylation and hydroxymethylation in Alzheimer's disease human brain[J]. Neurobiology of Aging, 2014, 35:1334-1344.

[13] AKHAVAN-NIAKI H, SAMADANI A A. DNA methylation and cancer development: molecular mechanism[J]. Cell Biochemistry & Biophysics, 2013, 67(2):501-513.

[14] ARAN D, SABATO S, HELLMAN A. DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes[J]. Genome Biology, 2013,14(3):2242-2254.

[15] RICKETTS C J, MORRIS M R, GENTLES D, et al. Methylation profiling and evaluation of demethylating therapy in renal cell carcinoma[J]. Clinical Epigenetics, 2013, 5(1):16-16.

[16] BAEK S J, YANG S, KANG T W, et al. MENT: Methylation and expression database of normal and tumor tissues [J]. Genes, 2013, 518(1): 194-200.