PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2021,48(10):1233~1242
www.pibb.ac.cn

基于多组学数据的早期宫颈鳞状细胞癌分类*

王晓曦　李晓琴**　曹阿成　侯智超　高　斌
（北京工业大学环境与生命学部，北京 100124）

摘要　从分子层面对泛癌进行研究已经得到了很大的进展，但是对宫颈鳞状细胞癌的分子分类研究仍然需要更多的探索. 为了找到宫颈鳞状细胞癌潜在的子类，本文提出了一个基于多维组学数据的癌症亚型分类分析流程. 通过统计学方法对癌症基因组图谱（The Cancer Genome Atlas，TCGA）宫颈鳞状细胞癌的mRNA表达数据、小分子核糖核酸（microRNA，miRNA）表达数据、DNA甲基化数据以及拷贝数变异数据4个维度包含的分子进行筛选，然后对筛选后的分类特征进行整合聚类，进一步筛选能够区分不同子类的关键分类特征，并使用这些关键分类特征建立宫颈鳞状细胞癌分类模型. 本研究为宫颈鳞状细胞癌分子层面子类的识别提供了分析流程，得到了两个临床生存水平具有显著性差异的宫颈鳞状细胞癌子类，并确定了8个宫颈鳞状细胞癌的关键分类特征. 本研究中识别的宫颈鳞状细胞癌子类和关键分类特征为宫颈鳞状细胞癌早期分类及分类标志物的鉴定提供了重要参考.

关键词　多基因组学，亚型分类，宫颈鳞状细胞癌，癌症早期
中图分类号　Q7，Q8　　**DOI**：10.16476/j.pibb.2020.0434

2018年的全球癌症统计表明，宫颈癌是女性发病率和死亡率均排名第四的癌症[1]. 宫颈鳞状细胞癌的组织学类型是鳞状细胞癌，是宫颈癌最常见的一种组织学类型，约占所有宫颈癌病例的90%[2]. 而鳞状细胞癌是一类上皮组织细胞、鳞状细胞产生病变的癌症，其病状、发病史、预后和治疗方法与其病发位置紧密相关[3]. 在过去的20多年里，宫颈癌的研究取得了重大进展，HPV疫苗的开发大大降低了宫颈癌的发病率和病死率[4]，然而被诊断为侵袭性和转移性的宫颈癌，依然没有较好的治疗方法[5]. 相关研究表明，宫颈癌早期诊断的患者生存率比晚期要高出5~10倍[6]. 因此，宫颈癌的早期诊断至关重要，而对分类特异性分子特征的鉴定，为探索早期癌症标志物奠定基础.

肿瘤的分类对抗癌药物的设计和选择起着重要的作用[7]. 肿瘤细胞起源可以影响但不能完全决定细胞分类[8]. 肿瘤多基因组数据的公开促进了分子层面的泛癌研究. 癌症基因组图谱（The Cancer Genome Atlas，TCGA）计划是由美国国家癌症研究所（National Cancer Institute，NCI）和国家人类基因组研究所（National Human Genome Research Institute，NHGRI）共同发起的癌症基因组计划，为关键基因组的突变生成了全面的多维图谱，提供了mRNA表达水平、小分子核糖核酸（microRNA，miRNA）表达水平、DNA甲基化、染色体变异、突变以及反相蛋白质阵列等多维基因组学数据，可以为宫颈鳞状细胞癌从分子角度进行分类的研究提供数据基础.

在肿瘤的分子分类方面，Hoadley等[8]对TCGA进行了全面的综合分子分析，包括大约10 000多个组织样本，33种类型的癌症，使用了多维基因组学数据进行聚类，找到了28种不同分子亚型. Bailey等[9]对33种癌症类型中超过9 000个肿瘤中的致癌驱动基因和突变进行了全面分析，突显了TCGA肿瘤样品中可用于临床的癌症驱动事件的普遍性. 以上研究让我们看到了结合细胞起源和分子分类学，得到更为稳定的肿瘤分类以及更显著

* 国家自然科学基金(61931013)和国家科技部重点研发项目(2017YFC0111104)资助.
** 通讯联系人.
Tel：15313254516，E-mail：lxq0811@ bjut.edu.cn
收稿日期：2020-12-09，接受日期：2021-03-19

的生物学特征和临床特性的可行性，但这些研究对结合特定部位和组织类型的肿瘤讨论不够深入. Berger等[10]对2 579例TCGA妇科和乳腺肿瘤进行分子分析，使用16个关键分子特征来识别5种预后类型，根据6种临床上可评估的特征提出了一个能够用于划分患者类型的决策树模型. 虽然该研究涉及了泛妇科癌，但是缺少对宫颈鳞状细胞癌的深入研究. Burk等[11]对宫颈癌全面基因组综合研究是迄今为止最大的宫颈癌分子层面的研究之一，确定了新的基因组和蛋白质组特征以及宫颈癌分类. 然而这些研究并没有针对宫颈癌的早期阶段进行讨论. 由于这些研究没有对用于分类的基因进行进一步的筛选，现有的研究结果可能会引入一些对聚类没有贡献的基因，从而扰乱结果.

在本研究中，通过分子分类学的方法对宫颈鳞状细胞癌早期样本进行分类. 首先对宫颈鳞状细胞癌I期样本的mRNA表达数据、miRNA表达数据、DNA甲基化数据和染色体变异数据进行分类特征筛选，然后使用筛选出的分子特征进行单个维度的层次聚类和综合所有维度数据的整合聚类，发现在临床生存上有显著性差异的子类，识别用于区分该子类的关键分类特征.

1 数据及数据预处理

从FireBrowse平台（http://firebrowse.org/）下载TCGA提供的宫颈鳞状细胞癌早期的mRNA表达数据、拷贝数变异数据、miRNA表达数据和DNA甲基化的数据. 从未治疗、原发性肿瘤样本和配对的癌旁样本中选择临床病理诊断为鳞状细胞癌或鳞状分化、病例临床信息的临床阶段为I期的样本. 对宫颈鳞状细胞癌期样本，筛选具有4个维度：mRNA表达数据、DNA甲基化数据、miRNA表达数据和拷贝数变异数据的样本，最终获得早期肿瘤样本120例，癌旁样本2例.

通过TCGA数据接口，下载转录组测序原始数据，涵盖52 421个基因；从FireBrowse下载DNA甲基化数据，该数据对每个基因选取β值具有最大标准偏差的探针，涵盖9 254个基因；从FireBrowse下载miRNA表达数据，从llumina hiseq/aga miRNA表达第3水平数据集中选择“RPM”（每百万miRNA前体读数），转换为$\log_2$的数据矩阵，包含543个特征数；从FireBrowse下载拷贝数变异数据，移除重复的拷贝数区域，对所有样本取所有唯一断点的并集，并移除过小的序列，最终得到包含689个特征的拷贝数变异数据.

对各个维度的数据，使用最大最小放缩法进行归一化将特征的值放缩到0~1之间：

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

经过数据获取和预处理后，得到4个维度的数据，每个维度特征数如表1.

Table 1 The number of features of the four-dimensional data preprocessed using the maximum and minimum normalization method

Dimension	Features
mRNA expression	52 421
DNA methylation	9 254
miRNA expression	543
Copy number varaition	689

2 方 法

本文通过对用于整合聚类的分类特征筛选和对整合聚类中起着关键作用的分类特征筛选，建立了基于多维组学数据的癌症亚型分类数据处理分析流程（图1）.

2.1 分类特征筛选方法

分子特征是将癌旁样本和肿瘤样本区分开的特征. 通过比较癌旁和肿瘤两组样本的均值分布区间的差异性筛选分子特征，均值分布区间如下：

$$\left[\mu - i \times \sigma, \mu + i \times \sigma\right] \tag{2}$$

其中，μ、σ分别为分子特征在肿瘤样本（或癌旁样本）的均值、标准差. 后续对于癌症样本分类的研究将只使用癌症样本.

分类特征是用于对癌症样本进行子类划分的分子特征. 使用变异系数的方法筛选出数据中离散程度较大的分子特征：

$$C_v = \frac{\sigma}{\mu} \tag{3}$$

其中μ、σ分别为分子特征在癌症样本中的均值、标准差. 变异系数越大说明该分子特征可能在癌症样本中的不同子类中水平差异越大.

2.2 宫颈鳞状细胞癌早期亚型的分类方法及关键分类特征筛选

在聚类前对各个维度的数据进行标准化，使用z分数标准化，计算公式为：

$$z = \frac{x - \mu}{\sigma} \tag{4}$$

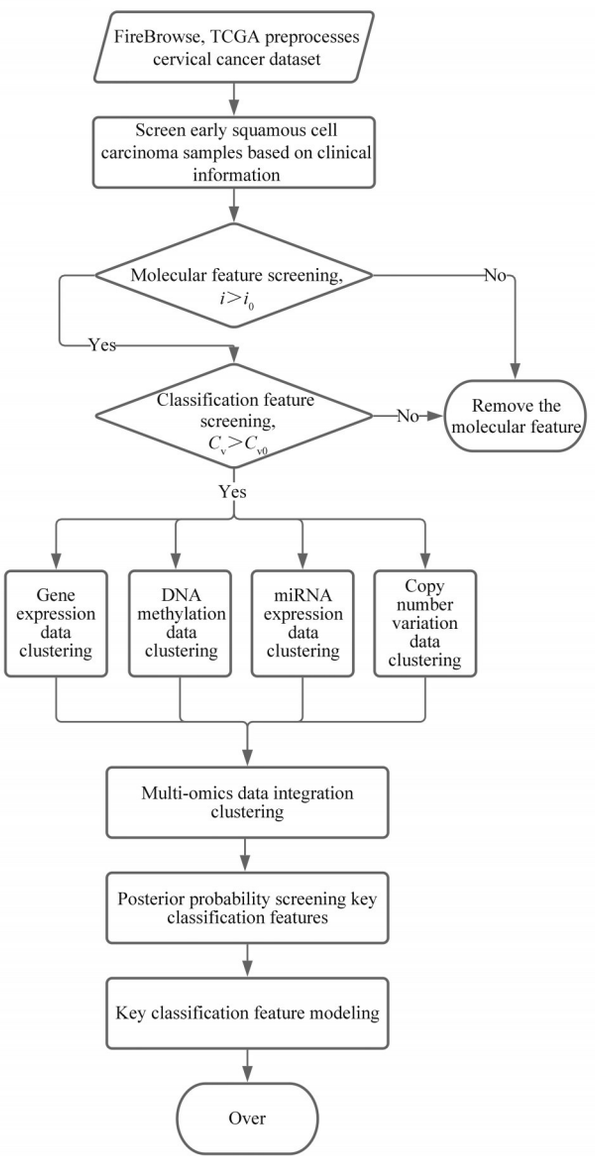

Fig. 1　Cancer subtype classification process based on multi-omics data

Preprocess the multi-omics data, screen molecular features and classification features, then perform integrated clustering and analysis on the multi-omics data after screening, and finally screen key classification features and verify modeling.

其中σ、μ分别为特征在癌症样本中的标准差、均值. 使用离差平方和法（Ward法）和欧几里得距离对每一维度的数据根据分类特征进行层次聚类. 层次聚类的结果以树状图的形式展现，可以直观的观测到类间的距离. 通过观测各个单维度组学数据的聚类结果，确定整合聚类的类数k.

对宫颈鳞癌早期样本，综合mRNA表达数据、miRNA表达数据、DNA甲基化数据和拷贝数变异数据的分类特征进行整合聚类. 选择基于贝叶斯的聚类算法[12]（iClusterBayes），聚类的类数为k.

根据整合聚类的结果，分析各个子类的组学特征，找到潜在的主导因素. 利用iClusterBayes方法计算各个维度的分类特征成为驱动因素的后验概率，可以用作选择驱动整体聚类基因组特征的标准. 通过调节后验概率的百分比，筛选分类特征获得关键分类特征，检验关键分类特征对分类的有效性，进一步验证癌症亚型分类流程的可行性.

3　结果及讨论

3.1　分类特征

对4个维度的数据使用均值分布区间差异性方法筛选能够区分癌症样本和非癌症样本的分子特征，然后用变异系数的筛选方法进行分类特征的筛选. 分子特征和分类特征的筛选是层次聚类和整合聚类的前置步骤，主要用于过滤掉在癌旁样本和肿瘤样本中特征水平无显著性差异的特征以及在肿瘤样本中离散程度低的特征，同时也可以降低数据维度，加快后续整合聚类模型收敛的速度. 考虑到TCGA数据集中的宫颈鳞状细胞癌早期样本的癌旁组织样本较少，所以对区分肿瘤样本和癌旁样本分子特征的筛选较为宽松. miRNA表达数据和拷贝数变异数据较少，均值分布区间的参数设置i_0和变异系数C_{v0}都设置为0.1. 分子特征和分类特征筛选参数的确定也参考了层次聚类的热图分布.

各维度的均值分布区间系数、变异系数及经过筛选后得到的4个维度数据的分子特征及分类特征数如表2.

Table 2　The parameter setting of molecular feature and classification feature screening and the number of features after screening

Dimension	i_0	Molecular features	C_{v0}	Classification features
mRNA expression	1.1	3 279	1.5	611
DNA methylation	0.5	6 740	1	765
miRNA expression	0.1	494	0.1	153
Copy number variation	0.1	689	0.1	120

3.2　分类结果及分析

3.2.1　基于分类特征的分类结果

图2a~d分别是基于mRNA表达数据、miRNA表达数据、DNA甲基化数据和拷贝数变异的分类特征数据进行层次聚类的早期宫颈鳞状细胞癌样本

热图，图中横轴为癌症样本，纵轴为分类特征，底部的标注为分类特征名称. mRNA表达数据可以发现仅存在靠上的一小支mRNA表达水平相对较高，很难通过mRNA表达数据的层次聚类结果对当前的样本进行分类（图2a）. 然而，miRNA表达数据显示，样本大体上也被分为了两支，靠上一支的部分分类特征出现了过表达（图2b）. DNA甲基化数据可以发现样本被聚为两大支，其中热图靠上部分的一支的样本，相对高甲基化的分类特征较多（图2c）. 拷贝数变异的分类特征数据显示，依据靠左的一簇分类特征，样本被分为两大支，分别对应拷贝数的扩增和拷贝数降低（图2d）.

由热图可见：基于mRNA表达、拷贝数变异和miRNA表达数据的宫颈鳞状细胞癌早期样本的聚类，不同分支对应的分类特征整体性特征并不明显；基于DNA甲基化数据的聚类，得到了相对甲基化水平较高的一支和相对甲基化水平较低的一支，且两支间的距离也比较大，这可能是由于筛选出的分类特征在早期样本中存在普遍的DNA甲基化现象，且分类特征在潜在的类别中存在差异.

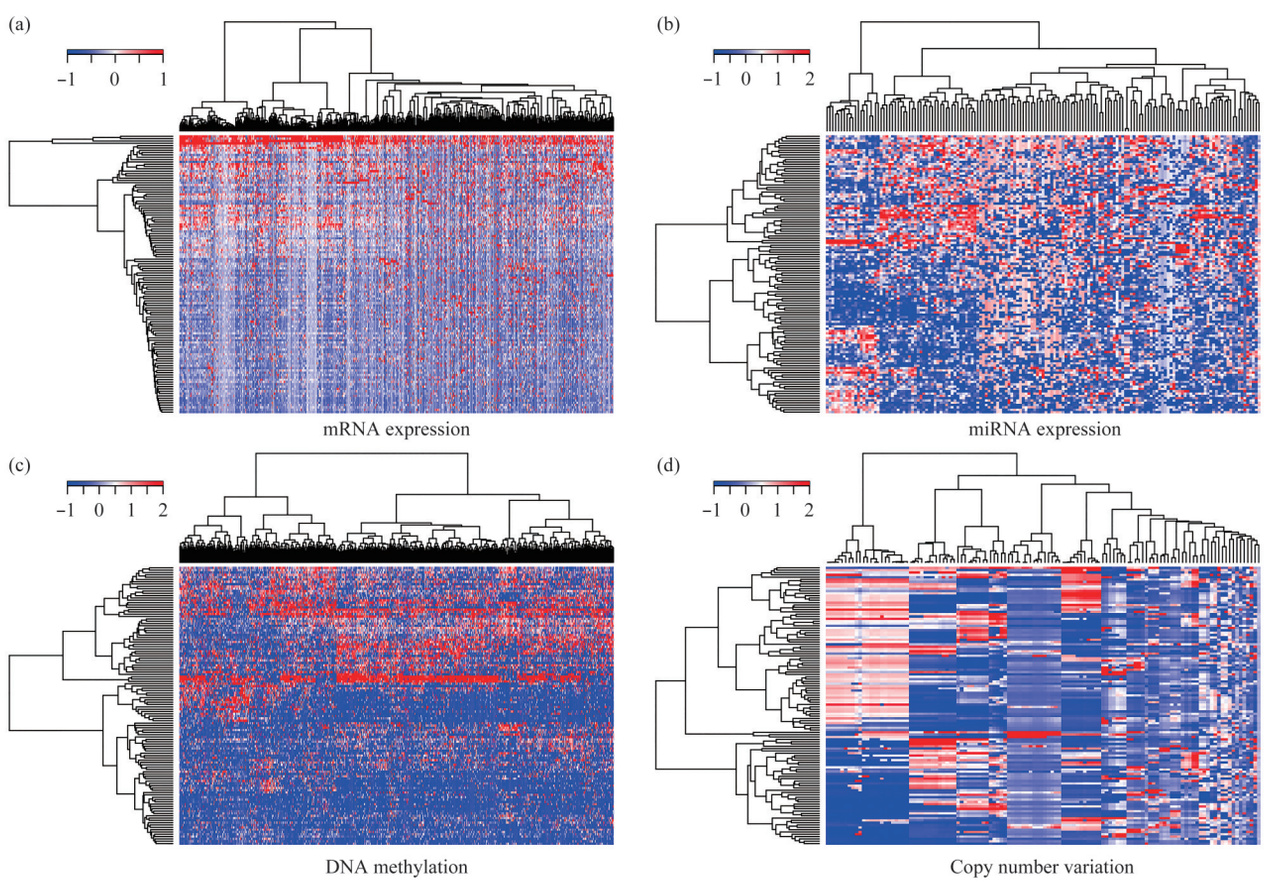

Fig. 2　Heat maps for hierarchical clustering of classification feature data in each dimension

The horizontal axis is the sample and the vertical axis is the classification feature.(a) mRNA expression, (b) miRNA expression, (c) DNA methylation, (d) copy number variation.

参考单组学数据的层次聚类结果，尝试将宫颈鳞状细胞癌早期样本聚为两类. 对mRNA表达数据、miRNA表达数据、DNA甲基化数据和拷贝数变异数据通过iClusterBayes的方法进行整合聚类（图3），横轴对应分类特征，纵轴对应样本. 将热图对应的2个分组分别命名为iC1（50例）、iC2（70例），代表宫颈鳞状细胞癌早期潜在的两个子类. 该热图主要揭示了早期宫颈鳞状细胞癌潜在的两个子类类别特征. mRNA表达数据全局表达量比正常水平略低，但在iC2中出现了少量特征相对高

表达. miRNA数据在iC2组中高表达的特征相对多一些，但两组的miRNA表达水平并没有显著的差异. 拷贝数变异数据的iC1、iC2没有发现较为明显的类别特征. 两个子类之间的组学差别主要体现在DNA甲基化水平上：iC1子类对应分类特征的高甲基化，iC2子类对应分类特征的低甲基化；iC1子类的高甲基化对应mRNA的更低表达，iC2子类的低甲基化则对应mRNA的少量高表达. 上述结果提示，宫颈鳞状细胞癌早期的分类可能主要由DNA甲基化主导.

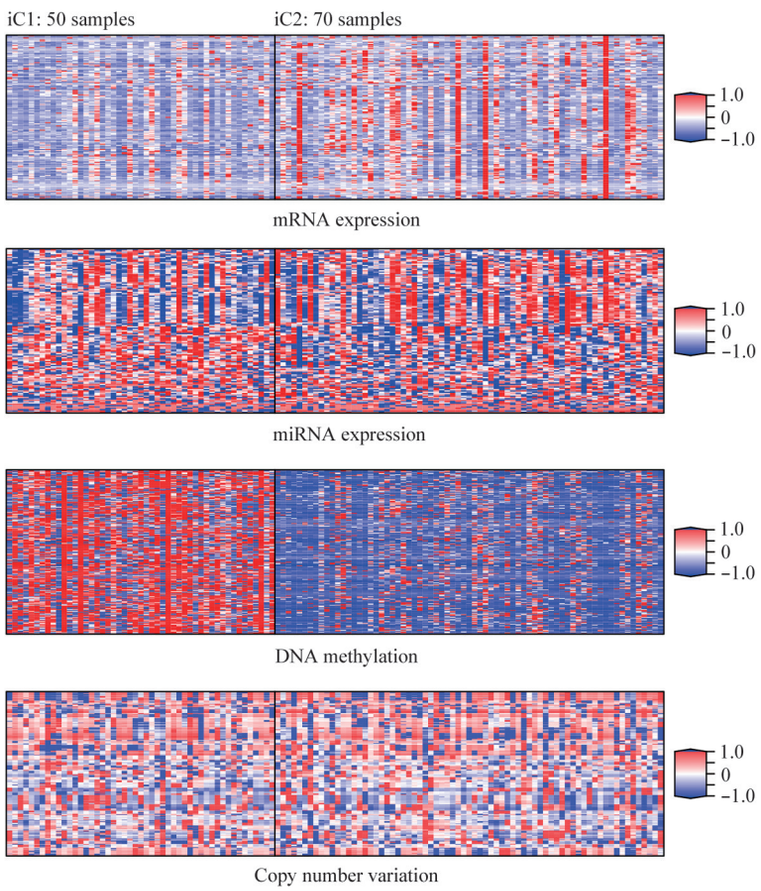

Fig. 3　The heat map of the multi−dimensional omics classification feature after screening is integrated and clustered by the iClusterBayes method

From top to bottom, there are heat maps generated by mRNA expression data, miRNA expression data, DNA methylation data, and copy number variation data based on the integrated clustering results. Among them, the horizontal axis is the classification feature, and the vertical axis is the sample. The vertical line in the middle of each sub-picture represents the dividing line of the category.

3.2.2　分类结果的稳定性

比较基于多维数据整合聚类的分类结果iC1、iC2和基于DNA甲基化数据层次聚类的分类结果hc1、hc2，绘制韦恩图（图4）. 从图中可以发现：iC1的50个样本和DNA甲基化聚类的第一类hc1有47个样本重合；iC2的70个样本和DNA甲基化聚类的第二类hc2有66个样本重合；整合聚类和DNA甲基化数据聚类的结果较为一致. 整合聚类结果与DNA甲基化聚类结果，是利用不同分类特征、使用不同聚类方法分别得到的，而分类结果却相对一致. 结果表明：分类结果的稳定性较好；宫颈鳞状细胞癌早期的分类主要受DNA甲基化影响和驱动. 进一步筛选DNA甲基化数据中的分类特征，可能得到能够区分两类的关键分类特征.

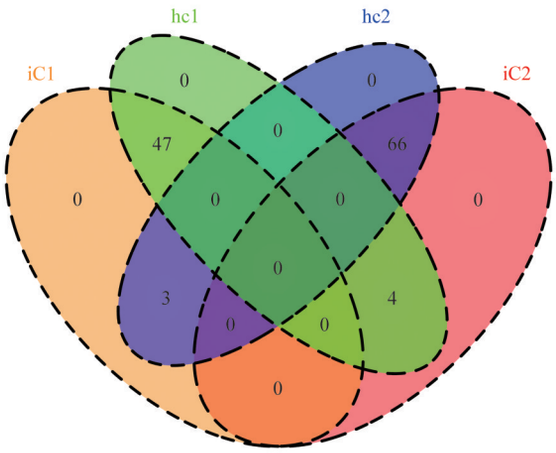

Fig. 4　DNA methylation clustering results and Venn diagram of integrated clustering results

Among them, hc1 and hc2 are the results of hierarchical clustering of DNA methylation, and iC1 and iC2 are the results of integrated clustering of multi-omics data.

3.2.3　亚型与临床结果相关性

对DNA甲基化数据、miRNA表达数据和拷贝数变异数据的层次聚类分类结果和整合聚类的分类结果进行生存分析（图5）. 前文分析了DNA甲基化层次聚类的结果和整合聚类的结果的一致性（**3.2.1**），所以这两组生存曲线也较为相似（图5b，d）. DNA甲基化层次聚类对应分类结果的生存曲线如图5b，hc1组和hc2组的生存情况存在显著性差异（P=0.002 1，秩检验）. 整合聚类的生存分析结果如图5d，iC1组和iC2组的生存情况也存在显著性差异（P=0.002 3，秩检验）. 说明DNA甲基化层次聚类和整合聚类都能得到预后具有差异的两个子类. 其他维度数据得到的层次聚类结果无法在生存情况上体现出差异. 后续需要探讨产生临床差异的关键因素，分析这些关键因素对预后的影响. 对于预后存在显著差异的两类，可能病状和治疗方法

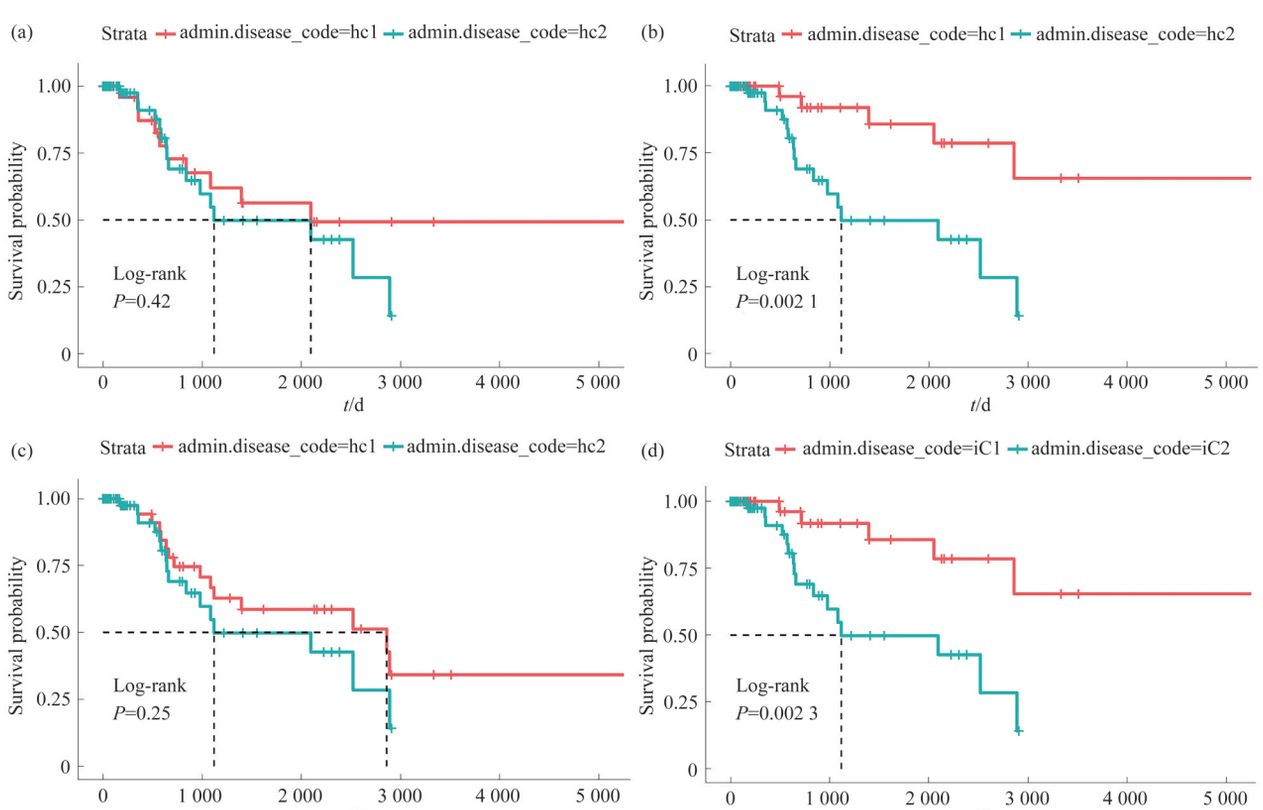

Fig. 5　Survival curve

After hierarchical clustering of each omics data, the survival curves of hc1 and hc2 and the survival curves of iC1 and iC2 obtained by integrating the omics data for integrated clustering: (a) Copy number variation hierarchical clustering, (b) DNA methylation hierarchical clustering, (c) miRNA expression hierarchical clustering, (d) Integrated clustering.

都存在不同.预后较差的一组需要更多的关注，以及开发出更有效的针对性疗法.

3.3　亚型的关键分类特征

3.3.1　关键分类特征

通过iClusterBayes算法计算各个维度的分类特征成为驱动因素的后验概率，进一步筛选关键分类特征.在单维度聚类分析中，DNA甲基化数据的层次聚类结果和整合聚类结果较为一致，均为两类，且两类对应的预后存在显著性差异.有研究发现肿瘤早期常常会出现DNA甲基化的改变[13]，因此DNA甲基化数据可能在整合聚类中对早期的宫颈鳞状细胞癌的分类起了主导作用.将后验概率占比前99%的8个DNA甲基化分类特征为关键分类特征，分别为BEGAIN、CD200、CHST13、FGF4、HMX3、LRRFIP1、PDE1B、PRDM12.

利用关键分类特征进行层次聚类并进行后续的分析（图6），图6a是基于关键分类特征的聚类结果，分别对层次聚类明显的两支做生存分析（图6b）.为了衡量分类结果的稳定性，对比关键分类特征的分类结果和整合聚类的分类结果，制作韦恩图（图6c），图中总计有14个样本分类存在差异，占总样本的11.7%，可见分类较为稳定.

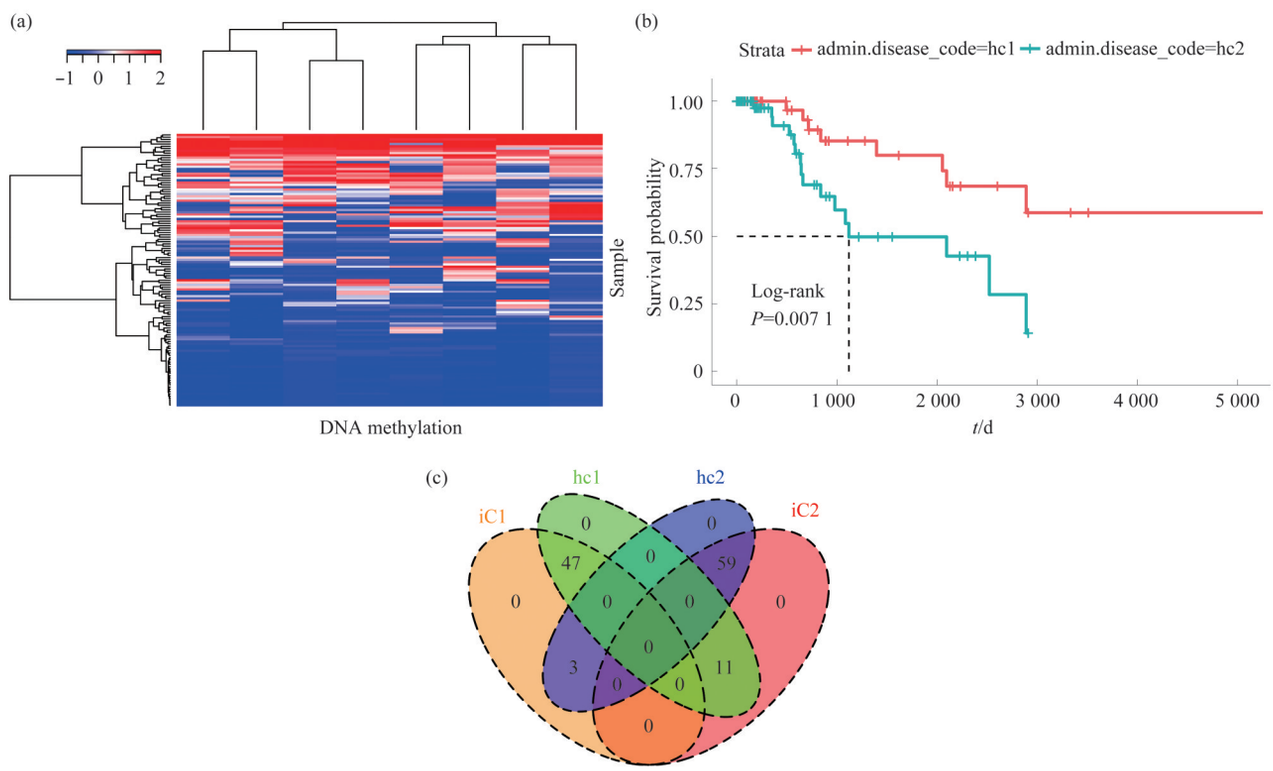

Fig. 6　Key classification features

(a) A heat map of hierarchical clustering using key classification features; (b) Two types of hc1 and hc2 obtained after hierarchical clustering of key classification features; (c) The Venn diagram of hc1 and hc2 obtained by hierarchical clustering of key classification features and iC1 and iC2 of the integrated clustering results.

关键分类特征可能在iC1和iC2这两个类别中的DNA甲基化水平存在差异，比较这些关键分类特征在整合聚类两个类里的DNA甲基化水平（图7）.发现上述8个基因在iC1和iC2这两类样本中DNA甲基化水平的分布存在明显的差异，而且预后较差的iC2 DNA甲基化水平相对较低，这说明这些关键分类特征可能起促癌作用，低甲基化促进了相关基因的表达.探索关键分类分子的功能，可能有助于理解和解释这两类具有显著性临床差异的宫颈鳞状细胞癌的发生机制.

我们对mRNA表达、miRNA表达、拷贝数变异数据进行了后验概率的筛选.mRNA表达数据筛选出的关键分类特征分是ADGRA2、HEPH、EMILIN1、ZNF660、MMP16、TCEAL7、MAGI2.AS3；miRNA表达数据筛选出的关键分类特征是hsa.mir.411、hsa.mir.758；拷贝数变异数据筛选出

的关键分类特征是chr3.145662790.162242059、chr11.104764995.112676385. T检验结果显示，上述不同组学数据对应的关键分类特征在iC1和iC2两类箱式图的四分位间距基本重合，分布无差异. DNA甲基化关键分类特征对应的基因与其他组学筛选出的关键分类特征无对应关系，且8个关键分类特征对应基因的甲基化数据与mRNA表达数据、拷贝数变异数据间相关系数的绝对值均低于0.3.

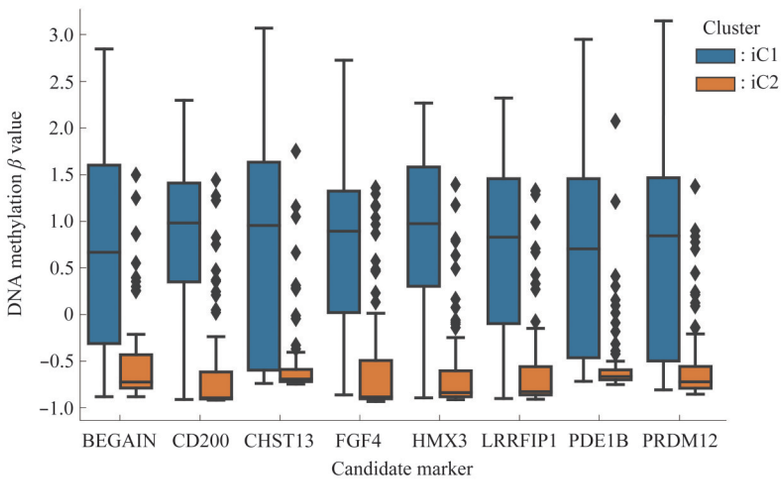

Fig. 7　Using Bayesian posterior probability to screen out the 8 key classification features in the integrated clustering of the two groups of iC1 and iC2 DNA methylation level comparison

关键分类特征均具有较为重要的生物学功能. CD200、FGF4、PRDM12是和肿瘤生长、转移相关的重要基因. CD200是一种耐受信号因子，该基因由多种细胞类型表达，包括B细胞、T细胞的子集、胸腺细胞、内皮细胞和神经元. 该基因编码的蛋白质在免疫抑制和抗肿瘤活性调节中起重要作用[14]. CD200可能在子宫内膜异位症患者的子宫内膜过度表达[15]. FGF4基因编码的蛋白质是成纤维细胞生长因子（fibroblast growth factors，FGF）家族的成员. FGF家族成员具有广泛的促有丝分裂和细胞存活活性，参与了肿瘤生长和侵袭的生物过程[16]. 相关研究表明，FGF4的扩增对食管鳞状细胞癌的预后存在影响[17]，FGF4的高度表达促进了乳腺癌细胞的转移[18]，在乳腺癌晚期存在FGF4的突变[19]. PRDM蛋白家族的成员几乎都与不同种癌症存在关联[20]. 说明本研究提出的基于多维组学数据的癌症亚型分类数据处理分析流程能够找到具有功能性的关键分类特征，为宫颈鳞状细胞癌早期的分类标志物提供了参考依据. PDE1B是许多重要生理过程的关键调节剂[21-22]. Ding等[23]在2020年通过重建宫颈癌的ceRNA生物学网络，发现PDE2A可能是CESC患者早期诊断和预后评估的生物标志. PDE2A和PDE1B都属于PDE家族，并在通路上存在着密切的关系.

CHST13编码的蛋白质属于磺基转移酶2家族，它位于高尔基体膜，并催化硫酸盐转移至软骨素中葡萄糖醛酸残基侧接β-1,4-连接的N-乙酰半乳糖胺（GalNAc）C4羟基. 硫酸软骨素是一种有硫酸化黏多糖链的糖蛋白，其可能为一种重要的生长调控因子，硫酸软骨素合成酶及相关基因的异常表达可能会对细胞的生长产生严重影响. 硫酸软骨素构成软骨中存在的主要蛋白聚糖，并分布在许多细胞和细胞外基质的表面[24-25]. 免疫细胞化学和免疫组织化学分析显示，BEGAIN在神经元的突触和细胞核中均有表达[26]. LRRFIP1是人体系统中的生物调节剂，它的相关途径包括疾病的先天免疫系统和FGFR2信号传导，它的调节异常会导致包含癌症在内的多种疾病[27].

基于多维组学数据的癌症亚型分类数据处理分析流程得到的关键分类特征，可能为宫颈鳞状细胞癌早期分类标志物的鉴定提供重要参考.

3.3.2　基于关键分类特征的亚型分类模型

关键分类特征在iC1和iC2中具有不同的DNA甲基化水平，为检验关键分子特征区分iC1和iC2两个子类的能力，本文建立基于关键分类特征的Logistic回归模型，并用于两个子类识别. 采用十折交叉验证，iC1和iC2子类样本的分类准确率为95.83%、敏感性为95.92%、特异性为95.77%. 说明

8个关键分类特征具有较好的分类属性，关键分类特征的选择是合理的. 基于关键分类特征的亚型分类模型的建立，也为宫颈鳞状细胞癌早期样本亚型识别提供了方法.

4 结　论

本文对宫颈鳞状细胞癌早期样本的4种维度数据，mRNA表达数据、miRNA表达数据、DNA甲基化数据和拷贝数变异数据，筛选了分类特征，然后通过iClusterBayes方法进行了整合聚类分析，鉴定了2种预后存在差异的亚型iC1和iC2. 通过计算特征成为驱动因素的后验概率，筛选了DNA甲基化数据的8个关键分类特征，构建了宫颈鳞状细胞癌的分类模型. 本研究为基于多组学数据分子层面的分类研究提供了分析流程和研究方法，并为宫颈鳞状细胞癌早期的分类标志物提供了参考依据.

参 考 文 献

[1] Bray F, Ferlay J, Soerjomataram I, *et al*. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin, 2018, **68**(6): 394-424

[2] PDQ® Adult Treatment Editorial Board. Cervical Cancer Treatment (PDQ®) - Health Professional Version[EB/OL]. Bethesda, MD: National Cancer Institute, 2016 [2021-12-04]. https://www.cancer.gov/types/cervical/hp/cervical-treatment-pdq

[3] Berman J J. Tumor taxonomy for the developmental lineage classification of neoplasms. BMC Cancer, 2004, **4**: 88

[4] Kane M A. Preventing cancer with vaccines: progress in the global control of cancer. Cancer Prevention Research, 2012, **5**(1): 24-29

[5] Uyar D, Rader J. Genomics of cervical cancer and the role of human Papillomavirus pathobiology. Clin Chem, 2014, **60**(1): 144-146

[6] Alberts B, Johnson A, Lewis J, *et al*. Molecular biology of the cell. New York: Garland Science, 2017: 237-298

[7] Berman J. Modern classification of neoplasms: reconciling differences between morphologic and molecular approaches. BMC Cancer, 2005, **5**: 100

[8] Hoadley K A, Yau C, Hinoue T, *et al*. Cell-of-origin patterns dominate the molecular classification of 10 000 tumors from 33 types of cancer. Cell, 2018, **173**(2): 291-304.e6

[9] Bailey M H, Tokheim C, Porta-Pardo E, *et al*. Comprehensive characterization of cancer driver genes and mutations. Cell, 2018, **174**(4): 1034-1035

[10] Berger A C, Korkut A, Kanchi R S, *et al*. A comprehensive Pan-cancer molecular study of gynecologic and breast cancers. Cancer Cell, 2018, **33**(4): 690-705.e9

[11] Burk R D, Chen Z, Saller C, *et al*. Integrated genomic and molecular characterization of cervical cancer. Nature, 2017, **543**(7645): 378-384

[12] Mo Q, Shen R, Guo C, *et al*. A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. Biostatistics, 2017, **19**(1): 71-86

[13] Ruike Y, Imanaka Y, Sato F, *et al*. Genome-wide analysis of aberrant methylation in human breast cancer cells using methyl-DNA immunoprecipitation combined with high-throughput sequencing. BMC Genomics, 2010, **11**: 137

[14] Rygiel T P, Rijkers E S K, Ruiter T D, *et al*. Lack of CD200 enhances pathological T cell responses during influenza infection. The Journal of Immunology, 2009, **183**(3): 1990-1996

[15] Clark D A, Dmetrichuk J M, Dhesy-Thind S, *et al*. Soluble CD200 in secretory phase endometriosis endometrial venules may explain endometriosis pathophysiology and provide a novel treatment target. Journal of Reproductive Immunology, 2018, **129**: 59-67

[16] Katoh M. FGFR inhibitors: effects on cancer cells, tumor microenvironment and whole-body homeostasis (Review). Int J Mol Med, 2016, **38**(1): 3-15

[17] Huang J, Song Q, Wang H, *et al*. Poor prognostic impact of FGF4 amplification in patients with esophageal squamous cell carcinoma. Human Pathology, 2018, **80**: 210-218

[18] Shi H, Li Y, Feng G, *et al*. The oncoprotein HBXIP up-regulates FGF4 through activating transcriptional factor Sp1 to promote the migration of breast cancer cells. Biochem Biophys Res Commun, 2016, **471**(1): 89-94

[19] Freitag C E, Mei P, Wei L, *et al*. Genetic alterations and their association with clinicopathologic characteristics in advanced breast carcinomas: focusing on clinically actionable genetic alterations. Human Pathology, 2020, **102**: 94-103

[20] Mzoughi S, Tan Y X, Low D, *et al*. The role of PRDMs in cancer: one family, two sides. Curr Opin Genet Dev, 2016, **36**: 83-91

[21] Yu J, Wolda S L, Frazier A, *et al*. Identification and characterisation of a human calmodulin-stimulated phosphodiesterase PDE1B1. Cell Signal, 1997, **9**(7): 519-529

[22] Bender A T, Beavo J A. PDE1B2 regulates cGMP and a subset of the phenotypic characteristics acquired upon macrophage differentiation from a monocyte. Proc Natl Acad Sci USA, 2006, **103**(2): 460-465

[23] Ding H, Xiong X X, Fan G L, *et al*. The new biomarker for cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC) based on public database mining. BioMed Research International, 2020, **2020**: 5478574

[24] Yorifuji K, Uemura Y, Horibata S, *et al*. CHST3 and CHST13 polymorphisms as predictors of bosentan-induced liver toxicity in Japanese patients with pulmonary arterial hypertension. Pharmacol Res, 2018, **135**: 259-264

[25] Zhou H, Li Y, Song X, *et al*. CHST11/13 regulate the metastasis and chemosensitivity of human hepatocellular carcinoma cells *via* mitogen-activated protein kinase pathway. Digestive Diseases and Sciences, 2016, **61**(7): 1972-1985

[26] Yao I, Iida J, Nishimura W, *et al*. Synaptic and nuclear localization of brain-enriched guanylate kinase-associated protein. Journal of Neuroscience, 2002, **22**(13): 5354-5364

[27] Takimoto M. Multidisciplinary roles of LRRFIP1/GCF2 in human biological systems and diseases. Cells, 2019, **8**(2): 108

Classification of Early Cervical Squamous Cell Carcinoma Based on Multi-omics Data*

WANG Xiao-Xi, LI Xiao-Qin**, CAO A-Cheng, HOU Zhi-Chao, GAO Bin

(*Faculty of Environment and Life of Beijing University of Technology*, *Beijing* 100124, China)

Abstract Molecular classification of cancer is the current frontier of cancer omics and tumor precision medicine. Although great progress has been made in molecular analysis of whole cancer, the molecular classification of cervical squamous cell carcinoma still needs more exploration. In order to find the potential subtypes of cervical squamous cell carcinoma, this paper proposed a data processing and analysis process based on the classification of cancer subtypes based on multi-omics data. Specifically, we analyzed mRNA, and microRNA (miRNA) expression data, as well as DNA methylation and copy number variation in cervical squamous cell carcinoma cases, using datasets obtained from The Cancer Genome Atlas (TCGA). Moreover, we identified molecules in each dimension, as well as integrated and clustered filtered classification features, and used them to distinguish different subtypes. The resulting key classification features were used to establish a classification model for cervical squamous cell carcinoma. The resulting key classification features were used to establish a classification model for cervical squamous cell carcinoma. Our results revealed two cervical squamous cell carcinoma subtypes, with significant differences across clinical survival levels, as well as 8 key classification features of cervical squamous cell carcinomas. These findings are expected to provide important references for early classification of cervical squamous cell carcinoma and identification of classification markers.

Key words multi-omics, subtype classification, cervical squamous cell carcinoma, early stage of cancer

DOI：10.16476/j.pibb.2020.0434

* This work was supported by grants from The National Natural Science Foundation of China（61931013）and the Key Research and Development Program（2017YFC0111104）.

** Corresponding author.

Tel：86-15313254516，E-mail：lxq0811@ bjut.edu.cn

Received：December 9，2020 Accepted：March 19，2021