DOI:10.12113/202002004

肺鳞状细胞癌发生的早期标志物及肿瘤预测模型

尚文慧,王晓曦,李晓琴*,高 斌
(北京工业大学 生命科学与生物工程学院,北京 100124)

摘 要:选取癌症基因组图谱数据库的肺鳞状细胞癌(Lung Squamous Cell Carcinoma,LUSC)样本作为数据集,在全基因组的水平上研究肺鳞状细胞癌病人从正常到发病Ⅰ期基因表达的变化,寻找与LUSC发病密切相关的早期标志物,并建立一种基于早期标志基因的肿瘤预测模型。方法 采用模式识别分类法和基因通路和功能分析相结合的筛选方法,对LUSC的早期标志物进行识别,并运用Fisher判别建立肿瘤预测模型。得到12个LUSC的早期标志物,分别是*CLDN*18, *CD*34, *ESAM*, *JAM*2, *CDH*5, *F*11, *F*8, *CFD*, *MRC*1, *MARCO*, *SFTPA*2和*SFTPA*1,机器学习建模后对LUSC早期癌症样本和正常肺组织样本的分类精度达到了98%以上。由基因*SFTPA*1和*ESAM*建立的LUSC早期肿瘤预测模型,对正常肺组织和LUSC肿瘤Ⅰ期样本的分类敏感性和特异性分别为99.18%和100%,并且独立验证集的分类准确率也在90%以上。结论 筛选出的12个早期分子标志物有望成为LUSC诊断的标志分子,并且建立的肿瘤预测模型具有极高的准确性,可以为LUSC的发生机理研究以及早期肿瘤预测提供帮助。

关键词:肺鳞状细胞癌;基因表达;肿瘤发生;早期标志物;诊断模型
中图分类号:Q7;Q81 **文献标志码**:A **文章编号**:1672-5565(2020)04-223-13

Early markers and tumor prediction models of lung squamous cell carcinoma

SHANG Wenhui, WANG Xiaoxi, LI Xiaoqin*, GAO Bin
(*College of Life Science and Bioengineering*, *Beijing University of Technology*, *Beijing* 100124, *China*)

Abstract: Lung squamous cell carcinoma(LUSC) samples selected from the cancer genome atlas (TCGA) database were used as dataset to investigate differences of gene expression in cancer patients from normal to stage I cancer at the whole genome-level. Early molecular markers of LUSC were explored, and a tumor prediction model based on early marker genes was established. The early markers of LUSC were identified by the combination of pattern recognition classification, gene pathway and functional analysis, and the prediction model was established by Fisher discriminant. According to the screening procedure, 12 early markers of LUSC were obtained, namely *CLDN*18, *CD*34, *ESAM*, *JAM*2, *CDH*5, *F*11, *F*8, *CFD*, *MRC*1, *MARCO*, *SFTPA*2, and *FTPA*1. Modeling by machine learning method, the classification accuracy rate of early cancer samples and normal lung tissue samples of LUSC was over 98%. Based on the selected early LUSC markers, the Fisher discriminant analysis method was used to establish a prediction model. The specificity and sensitivity of the LUSC early tumor prediction model established on the basis of *SFTPA*1 and *ESAM* for normal lung tissue and stage I cancer samples were 100% and 99.18%, respectively. The classification accuracy of the independent validation set was more than 90%. The 12 early molecular markers are expected to be the marker molecules for the diagnosis of LUSC, and the established tumor prediction model has high accuracy, which can be helpful for the study of the pathogenesis of LUSC and early tumor prediction.

Keywords: Lung squamous cell carcinoma; Gene expression; Tumorigenesis; Early markers; Prediction model

---

收稿日期:2020-02-10;**修回日期**:2020-04-20.
基金项目:国家自然科学基金项目(No.11572014); 国家科技部重点研发项目(No.2017YFC0111104).
作者简介:尚文慧,女,硕士研究生,研究方向:生物信息学. E-mail:1272005180@qq.com.
*通信作者:李晓琴,女,教授,研究方向:生物信息学.E-mail:lxq0811@bjut.edu.cn.

肺癌是常见的恶性肿瘤之一,根据世界卫生组织统计,肺癌的发病率和死亡率均居全球恶性肿瘤的首位。肺癌按组织学分类一般分为小细胞肺癌(Small cell lung cancer, SCLC)和非小细胞肺癌(non-small cell lung cancer, NSCLC),NSCLC 约占肺癌的 80%,肺鳞状细胞癌(Lung squamous cell carcinoma, LUSC)是 NSCLC 的主要病理类型之一,导致全世界每年约有 40 万患者死亡[1]。然而,肺癌的发病机制迄今尚不明确,且起病隐匿。早期的肺癌一般没有明显症状,70%的患者临床发现时已是晚期,失去了临床治疗的最佳时机。肺癌的 5 年的生存率只有 8.9%~15%,而Ⅰ期的肺癌术后 5 年生存率高达 80%,这表明患者存活率与疾病的诊断阶段密切相关。因此,早期诊断对于癌症的治疗和预后都起着关键作用。

近年来,随着分子生物学研究的不断深入,对癌症的发病机制的认识也不断加深,分子基因标志物在癌症临床诊断和治疗中的作用日益受到关注。Liu 等[2]人发现 *TRIM*28 基因可以作为早期非小细胞肺癌转移和预后的标志物,非小细胞肺癌患者中 *TRIM*28 的总阳性率为 30.4%(138 例中 42 例),早期患者总阳性率为 29.9%(97 例中 29 例)。Tseng 等[3]研究 *SLIT*2 在肺癌进展中的基因表达水平,结果表明 *SLIT*2 可以抑制肺癌进展,并认为其可能为肺癌治疗及预后的潜在“治疗靶标”。Yu 等[4]也报道了 PEBP4 作为分子标志物与肺鳞状细胞癌的发生发展、浸润转移以及分化有关,LUSC 患者中 PEBP4 的总阳性率达到了 93.4%,然而 PEBP4 依赖肿瘤阶段,早期(Ⅰ期、Ⅱ期)患者 PEBP4 表达明显低于晚期(Ⅲ期、Ⅳ期)患者($p < 0.05$)。由于单一的基因标志物对 LUSC 的检出率不高,并且早期 LUSC 标志物较少,因此探索新的可靠的分子标志物对 LUSC 的诊断和治疗意义重大。

基因表达水平与癌症的发生密切相关。众所周知,产生癌变的因素有很多,包括基因突变,抑癌基因的功能丧失,原癌基因的激活,以及其他与癌症相关的因素。基因表达数据代表了每个基因的即时表达数据,从这些数据中能够挖掘到有用的信息,发现与癌症相关的基因标志物。癌症基因组图谱计划数据库(The cancer genome atlas, TCGA)和基因表达组数据库(Gene expression omnibus, GEO)为研究人员提供了大量的基因表达数据,如何克服高维小样本的特点进行特征基因的挖掘是生物信息学研究的一个难点,现有研究通常运用统计学方法来挖掘与肿瘤相关的基因。李建更等[5]提出一种多步降维法,该方法运用基因表达差异显著性分析方法(SAM)、偏最小二乘 VIP 系数法(PLS)和基于巴氏距离的顺序前向搜索方法(BD-SFS),最终提取到 20 个能将胃癌亚型有效分开的特征基因,利用支持向量机作为分类模型,准确率达到 89.43%。Zhang 等[6]通过迭代降维递归法筛选出 67 个特征基因,对 LUSC 癌症Ⅰ期、Ⅱ期和Ⅲ期样本进行分类,准确率达到 86.3%。Feng 等[7]因表达值筛选得到一组 188 个基因组成的基因团,能够较准确地区分乳腺原发癌与癌旁的正常乳腺组织。Lau 等[8]使用基因表达数据为患者总体生存率建立三基因(*STX*1*A*、*HIF*1*A*、*CCR*7)分类器,该分类器能够对非小细胞肺癌Ⅰ期和Ⅱ期患者进行分类,并辅助改善组织学对肿瘤阶段的预测能力。以上研究表明利用生物信息学方法筛选癌症相关分子基因标志物的可行性,且筛选得到的基因具有较好的分类效果。

关于肿瘤发生的候选生物标志物筛选我们作了一些工作[9-10],也取得了比较好的结果。上述工作主要基于基因表达的差异度及其对分类的贡献大小来筛选,在此基础上发现一些生物学功能明确且生物学过程清楚的分子标志物并建立预测模型是本文的重点。通过统计学并结合基因的生物学通路分析筛选早期 LUSC 的分子标志物的方法,为肿瘤标志物的筛选提供了新的视角,筛选得到的标志物可能成为 LUSC 的诊断及治疗靶点,有助于 LUSC 的分子机制的研究,建立的肿瘤预测模型能够提高 LUSC 的与预测准确率,为科学研究和临床诊断提供了一种新途径。

1 数据与预处理

1.1 数据

1.1.1 TCGA data

训练集数据来源于癌症基因组图谱(The Cancer Genome Atlas, TCGA)公共数据库,下载的数据包括癌症基因表达谱数据和临床信息数据。使用 Python 语言编写程序将病人的表达谱和临床信息进行整合,挑选出正常组织和癌症Ⅰ期的样本表达谱数据,同时为了便于后续的研究,删去在所有样本中表达值为空的基因。

最终收集到肺鳞状细胞癌样本 295 例(LUSC,49 正常人,245 癌症Ⅰ期)、甲状腺癌样本 342 例(THCA,56 正常人,286 癌症Ⅰ期)、肝细胞癌样本 222 例(LIHC,49 正常人,173 癌症Ⅰ期)和肾透明细胞癌样本 345 例(KIRC,72 正常人,272 癌症Ⅰ期)。

1.1.2 GEO data

从 GEO 数据库下载肺癌的基因表达数据(数据

集为 GSE11969),作为独立测试集对模型进行检验。仅保留信息完整的 LUSC 肿瘤 I 期和正常肺组织样本的基因表达数据,共得到样本数 22 例,其中癌症 I 期样本有 17 例,癌旁组织样本 5 例。

1.2　预处理

由于基因的表达数据相差较大,为了方便后期建模且训练收敛迅速,需要对基因表达谱数据进行归一化处理,要求其区间为[0,1],取值公式为:

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

其中:$x_{\min}$ 为该基因在所有样本表达值的最小值;$x_{\max}$ 为该基因在所有样本表达值的最大值。

2　方　法

2.1　肺鳞状细胞癌早期特征基因的筛选方法

本文涉及的分类问题为常见的二分类模型,但是由于特征的维度远远大于样本数,特征之间的关联关系相对复杂、关联关系间依赖性影响等问题,使得学习产生了诸多问题,比如:分析数据、训练模型时间长,数据量大导致“维度灾难”,使得模型过于复杂等等。为了克服这些不利因素的影响,提高特征识别的准确率,需要对数据集的特征基因进行筛选。本文在运用统计学的基础上,结合基因的生物学功能,建立一套分子标志物的筛选流程。

(1) 相关性筛选　相关性筛选可以计算出两个指标之间的相关系数,相关系数越大代表两个指标信息之间的相关性越强。在这里,两个指标分别代表着基因的表达值于癌症分期(正常肺组织和癌症 I 期)的相关性。本文采用斯皮尔曼公式计算相关性,设定相关系数 r,保留相关系数 r 大于 0.5 的基因作为候选基因。

(2) t 检验筛选　t 检验过程是对两个癌症分期(癌症 I 期和正常)均数差别的显著性进行检验。对于筛选出的特征基因,满足齐性分布的采用双总体 t 检验进行筛选,保留对癌症分期具有显著性差异的特征基因,不满足齐性分布的基因采用非参检验中的 Cruskal-Wallis 秩和检验进行筛选,同样保留对癌症分期具有显著性差异的特征基因。

(3) 置信区间筛选　置信区间是一种常用的区间估计方法,根据基因的表达在癌症 I 期和正常肺组织中的差异性筛选基因,通过置信区间是否重合筛选出一组具有明显差异表达的基因,作为肺鳞状细胞癌差异基因子集 1。其中,置信区间的筛选 i 值设为 1.0。

(4) 弹性网络筛选　弹性网络可以克服噪声和变量相关性的影响,原理是将分类贡献较小的自变量的系数降为零。并且,当特征之间存在较强的共线性时,弹性网络可以标记出所有的变量,从而形成一个特定的基因组合。运用弹性网络($\alpha = 0.5$)的方法迭代筛选出另一组最优肺鳞状细胞癌差异基因子集 2。

(5) KEGG 通路富集筛选　KEGG 通路分析可以确定不同样本间差异基因所参与的最主要代谢途径和信号转导途径。通过对差异基因做通路分析和生物学功能分析可以辨认 LUSC 的重要通路及其重要基因。采用 DAVID 在线分析平台对上述的两组特征基因集(LUSC 特征基因集 1 和 LUSC 特征基因集 2)分别进行 KEGG 通路分析,找出两组特征基因参与的共同通路,并进一步筛选出通路内的共有基因,得到候选基因集。具体流程图如图 1 所示。

为了验证筛选出的基因具有足够高的癌症分类能力,我们需要对候选基因集进行模式识别模型预测。对候选基因集,采用支持向量机(Support Vector Machine, SVM)、随机森林(Random Forest , RF)、人工神经网络(Artificial Neural Network, ANN)多种分类模型进行建模预测,并利用 10 折交叉验证法对模型参数进行优化。

用混淆矩阵作为模型的评判标准,混淆矩阵的概念图如图 2 所示。其中,Positive 是肺鳞状细胞癌癌症 I 期样本,Negative 是正常肺组织样本。由此可推出,TP 是指实际上是癌症 I 期样本并且也推断是癌症 I 期样本的样本个数;TN 是指实际上是正常肺组织样本,预测值也为正常肺组织样本的样本个数;FN 是指实际上是癌症 I 期样本,但预测为正常肺组织样本的样本个数;FP 是指将正常肺组织样本预测为癌症 I 期样本的样本个数。

通过计算总体准确率(Accuracy, ACC)以及敏感度(Sensitivity, SEN)、特异度(Specificity, SPE)、马修斯相关系数(Matthews Correlation Coefficient, MCC)指标来评价模型的分类结果。其中,敏感性指的是肺鳞状细胞癌癌症 I 期样本的分类准确率,特异性指的是正常肺组织样本的分类准确率,而马修斯相关系数指标考虑到了 TP、FP、TN 和 FN,对于不平衡的数据集也可以衡量。

ACC,SEN,SPE 与 MCC 的定义如下:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}} \tag{2}$$

$$\mathrm{SEN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

$$\mathrm{SPE} = \frac{\mathrm{TN}}{\mathrm{FP} + \mathrm{TN}} \tag{4}$$

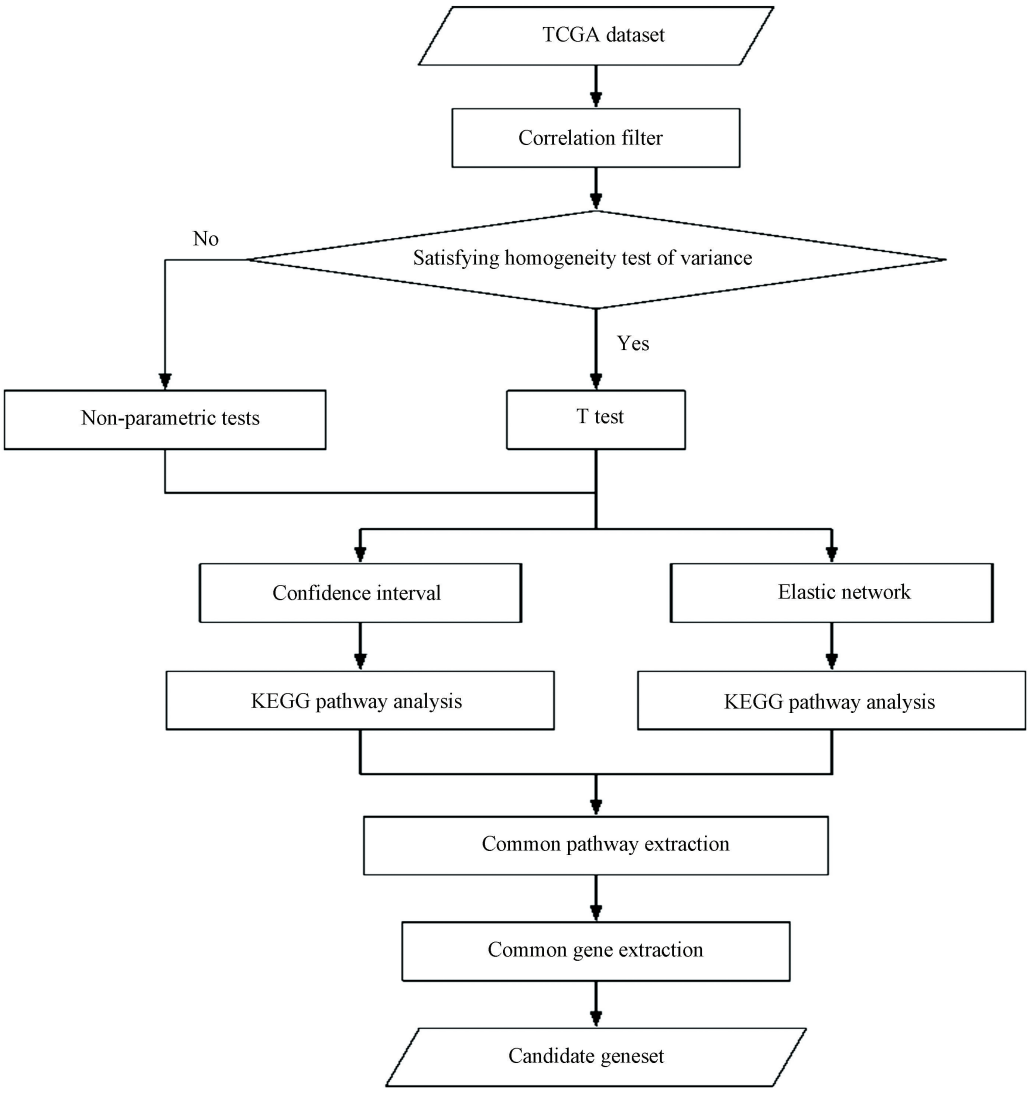

图 1 肺鳞状细胞癌分子标志物的筛选流程图

Fig.1 Flow chart of molecular markers identification for LUSC

混淆矩阵（Confusion matrix）		预测值（Predicted value）	
		Positive	Negative
观测值（Observed valuc）	Positive	TP (True positive)	FN (False negative)
	Negative	TP (False positive)	TN (True negative)

图 2 混淆矩阵概念图

Fig.2 Confuse matrix concept diagrams

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (5)$$

2.2 Fisher 判别建模

Fisher 判别分析是多元统计分析判别归属的一种方法，其基本思想是将高维数据点投影到低维空间（一条直线）上，找到一个投影轴使得样本投影到该空间后能在保证方差最小的情况下，将不同类的样本更好的分开。它能根据已有类别的若干样本的数据信息，总结出分类的规律性，建立判别公式和判别准则，当遇到新的样本时，能判定该样本所属的类别。

本节的目的是在筛选出的分子标志物的基础上，建立基于 Fisher 判别分析的分类预测模型。首先，以候选基因集中的每个基因作为单一变量建立模型；然后，对 12 个分子标志物进行两两配对建模；最后，对所有基因按照准确率的大小进行排列，依次迭代建模。以模型的准确率（Accuracy，ACC）以及敏感度（Sensitivity，SEN）、特异度（Specificity，SPE）、马修斯相关系数（Matthews Correlation Coefficient，MCC）作为指标来评价模型的分类结

果,通过多个模型的对比,选择表现最好的模型作为最终的分类预测模型。

3　结果与讨论

3.1　分子标志物的提取结果与讨论

3.1.1　分子标志物提取结果

在运用统计学方法的基础上综合分析基因的生物学功能,通过差异基因在通路上的富集,筛选得到 3 条共有通路和 12 个共有基因,如表 1 所示,筛选出的共有基因作为候选分子标志物组成候选基因集。

ROC 曲线可以筛查出对癌症的识别能力较高的基因,曲线下的面积越大,说明该基因的诊断效能越大,故保留 ROC 曲线下面积大于 0.9 的基因。对所有的候选分子标志物进行 ROC 曲线分析,发现所有候选分子标志物的 AUC 的值都在 0.98 以上,远远大于 0.9,说明候选基因对癌症的诊断有意义,所以最终保留候选基因集中的 12 个基因作为分子标志物,分别是 *CLDN*18, *CD*34, *ESAM*, *JAM*2, *CDH5*, *F*11, *F*8, *CFD*, *MRC*1, *MARCO*, *SFTPA*2, *SFTPA*1。

表 1　肺鳞状细胞癌的 KEGG 通路分析结果

Table 1　KEGG pathway analysis of LUSC

Term	Count	Genes
Cell adhesion molecules	5	*CLDN*18, *CD*34, *ESAM*, *JAM*2, *CDH5*
Complement and coagulation cascades	3	*F*11, *F*8, *CFD*
Phagosome	4	*MRC*1, *MARCO*, *SFTPA*2, *SFTPA*1

3.1.2. 相关通路及分子标志物的功能性分析

如前所述,经过 KEGG 通路分析后共富集到 3 条共有通路,分别是细胞粘附分子通路、补体通路和吞噬体通路,相应基因标志物在通路中的相关位置用红色星星加以标注。筛选得到 LUSC 的分子标志物共有 12 个,分别是 *CLDN*18, *CD*34, *ESAM*, *JAM*2, *CDH5*, *F*11, *F*8, *CFD*, *MRC*1, *MARCO*, *SFTPA*2, *SFTPA*1。

图 3 是分子标志物在 LUSC 癌症 I 期和正常组织中的基因表达水平,由图可以看出,这 12 个基因在癌症 I 期中表达量都低于正常组织组织中的表达量,说明这些基因的表达在 LUSC 早期的发生发展中受到了抑制。

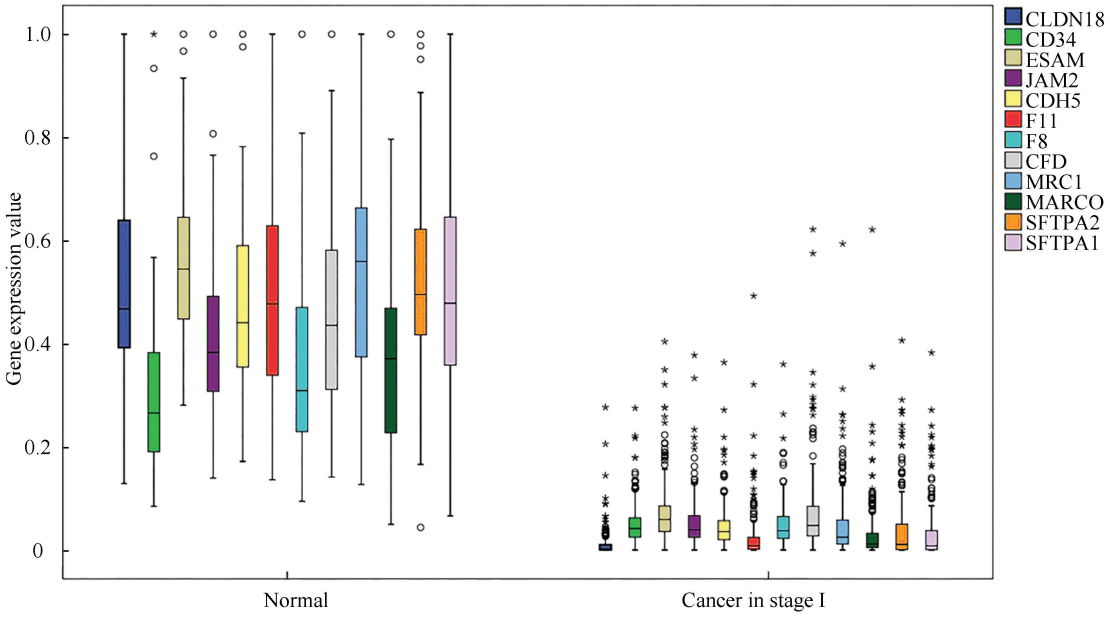

图 3　在肺鳞状细胞癌癌旁组织和癌症 I 期中基因表达水平箱线图

Fig.3　Box chart of gene expression levels in normal tissue and stage I cancer

(1)细胞粘附分子通路图如图 4 所示,12 个基因标志物中的 *CLDN*18、*CD*34、*ESAM*、*JAM*2、*CDH5* 等基因属于该通路且处于重要位置,这些基因编码的蛋白作为细胞粘附分子在通路中发挥作用。细胞黏附分子是众多介导细胞间或细胞与细胞外基质间相互接触和结合的膜表面糖蛋白分子的统称,以受

体–配体结合的形式发挥作用,使细胞与细胞间发生粘附,细胞间黏附作用密切参与体内免疫应答、炎症发生、凝血、肿瘤转移以及创伤愈合等一系列重要生理病理过程。由图 2 可知,基因 *CLDN*18、*CD*34、*ESAM*、*JAM*2、*CDH*5 在癌组织中的表达下调,这会导致细胞膜的通透性增加,为肿瘤细胞的侵袭和转移提供了机会。*CLDN*18 编码的蛋白质位于上皮细胞和内皮细胞,是紧密连接的重要结构成分[11],紧密连接结构的丧失可导致肿瘤细胞的侵袭和转移[12]。有文献报道,*CLDN*18 与胃癌的发生有关[13-14]。*CD*34 编码的Ⅰ型跨膜糖蛋白作为一种黏附分子,在乳腺肿瘤中出现了异常表达,与乳腺肿瘤从良性到恶性发展过程中起重要作用[15]。*ESAM* 表达于内皮细胞,在血管紧密连接中含量丰富,该基因的缺失可使内皮细胞的通透性增加,即 *ESAM* 的表达同细胞通透性呈负相关,表明该基因与肿瘤细胞的侵袭和转移有重要联系[16]。*JAM*2 在上皮细胞和内皮细胞的细胞连接处以及红细胞、白细胞和血小板的表面富集,Kok-Sin 等[17]发现 *JAM*2 基因在结直肠癌中高甲基化、低表达,与肿瘤的发生发展有关。*CDH*5 编码的蛋白在内皮细胞粘附连接的组装和维持中发挥作用,*CDH*5 在多种肿瘤中异常表达,包括侵袭性乳腺癌、非转移性肾细胞癌、侵袭性黑色素瘤和骨肉瘤等,与癌症的发生发展密切相关[18,21]。

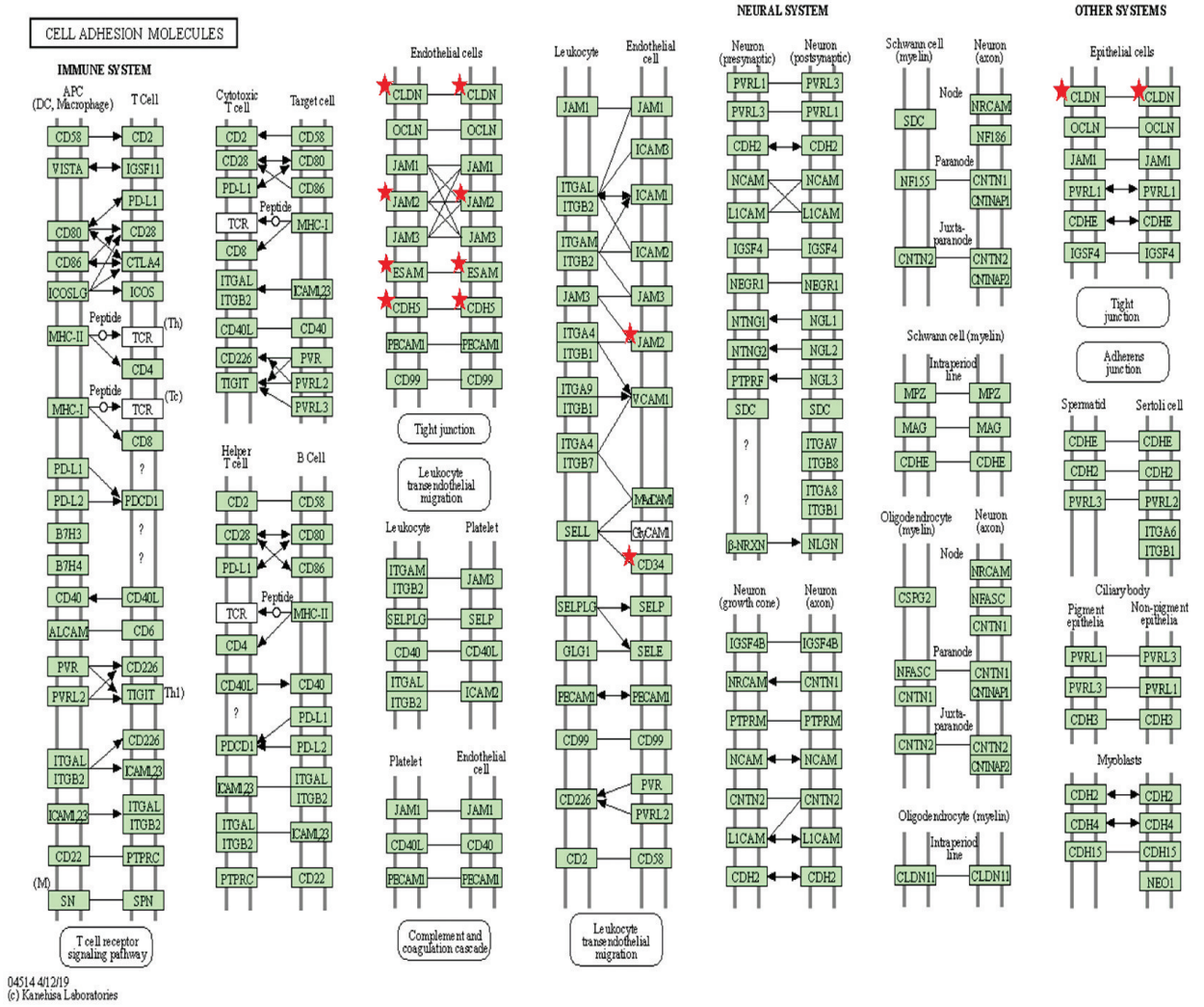

图 4 *CLDN*18、*CD*34、*ESAM*、*JAM*2 和 *CDH*5 在细胞粘附分子中的位置

Fig.4 Positions of *CLDN*18, *CD*34, *ESAM*, *JAM*2, and *CDH*5 in cell adhesion molecules

(2)补体系统通路图如图 5 所示,基因 *F*11、*F*8 和 *CFD* 在该通路中处于重要位置。补体系统又分为两条通路,分别是凝血级联和补体级联,它们相互独立又相互联系。其中,凝血级联是机体形成血块以防止失血的过程,*F*11 和 *F*8 编码的蛋白是凝血级联中的凝血因子,参与了凝血的内在途径,在肿瘤的微环境中,微小血栓的形成会限制血液供应从而抑制肿瘤细胞的生长,进而限制肿瘤细胞向周围组织浸润及转移[22];补

体级联中起重要作用的是补体,补体是一种存在于血清、组织液和细胞膜表面的一组经活化后具有酶活性的蛋白质,对机体的防御功能、免疫系统功能的调节以及免疫病理过程都发挥重要作用,*CFD* 编码的补体因子存在于补体系统的旁路途径中,具有级联放大作用,在文献[23]中发现,肝细胞癌患者的血清补体含量低于健康样本,可能的原因为肿瘤细胞表达补体的抑制剂,从而抑制补体的激活,阻止了补体的级联放大效应并大量消耗补体,导致肿瘤对机体免疫监督的逃逸,促进了肿瘤的发生与进展[24]。

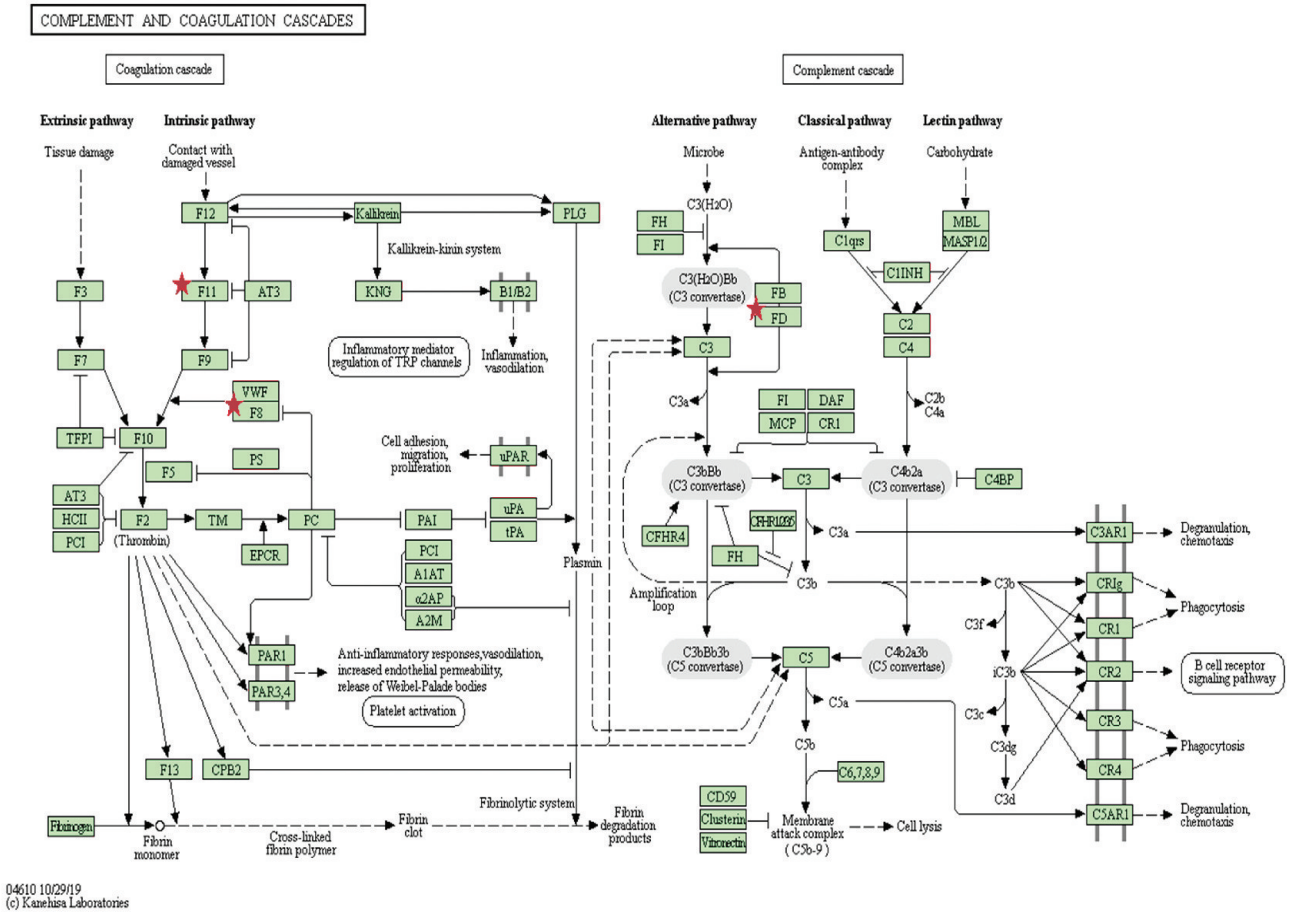

图 5　F11、F8 和 *CFD* 在补体通路中的位置

Fig.5　Positions of *F11*, *F8*, and *CFD* in complement and coagulation cascades

(3) 吞噬通路图如图 6 所示,基因 *MRC*1、*MARCO*、*SFTPA*2 和 *SFTPA*1 在该通路中发挥着重要作用。吞噬体是巨噬细胞参与组织重塑、清除凋亡细胞、抑制细胞内病原体扩散的天然能力的关键细胞器,是一种在免疫过程中常见的细胞结构。巨噬细胞的吞噬作用是细胞内化外来的大颗粒物质的过程,在先天免疫和适应性免疫中非常重要,有助于我们对抗传染病的能力。巨噬细胞在不同的环境下会发生不同性质的活化,而活化后的巨噬细胞具有不同的免疫功能。通常巨噬细胞有 3 种活化形式,分别是经典活化的巨噬细胞(M1 型巨噬细胞)、替代性活化的巨噬细胞(M2 型巨噬细胞)和 II 型活化的巨噬细胞。其中,M1 型和 M2 型巨噬细胞与肿瘤的发生、发展和转移密切相关,M1 型巨噬细胞能够分泌一氧化氮(NO)等杀伤分子,可以促进炎症的发展,加速感染的病原体和肿瘤细胞的凋亡;而 M2 型巨噬细胞被诱导剂活化后,使得甘露糖受体和清道夫受体表达上调,减少 NO 的分泌,从而降低了杀伤细胞内病原体的能力,对肿瘤细胞的增值和侵袭起促进作用[25]。基因 MRC1 编码的甘露糖受体和 *MARCO* 编码的清道夫受体是肺泡巨噬细胞的膜受体,其在早期 *LUSC* 中的低表达可能预示着肺泡巨噬细胞活化为 M1 型巨噬细胞,从而对肿瘤细胞的发展起抑制作用,并在抗原识别、内毒素清除、细胞粘附和调控炎性递质分泌等机体防御反应中起重要作用[26]。*SFTPA*1 和 *SFTPA*2 是两个同源基因,属 C 型胶凝素超家族(Collectins)成员,由肺泡 II 型上皮细胞分泌,作为肺内原始的内源性的免疫调节剂,具有维持肺部稳定、调节局部免疫和炎症反应以及参与局部防御等功能[27-28],且由于 *SFTPA* 具有肺

部特异性,因此其有望用于肺部疾病的治疗及作为新的肺疾病生物学标志物[29]。*MRC*1 和 *MARCO* 作为肺泡巨噬细胞表面的模式识别受体,可以识别病原体从而引发吞噬运动,而 *SFTPA*1 和 *SFTPA*2 编码的 *SFTPA* 能与许多不同的病原体包括病毒、真菌、细菌结合,加速它们被肺泡巨噬细胞吞噬和消灭,这在肺的原发性防御免疫系统中发挥着重要作用。

综上所述,通过对分子标志物的相关通路以及基因功能性分析,发现 *CLDN*18,*CD*34,*ESAM*,*JAM*2,*CDH*5,*F*11,*F*8,*CFD*,*MRC*1,*MARCO*,*SFTPA*2 和 *SFTPA*1 这 12 个分子标志物与 LUSC 的发生发展密切相关,已有文献也验证了其在癌症中的重要性,这也表明本文的筛选方法是可靠的。

3.1.3. 评估 LUSC 的分子标志物

(1)对 LUSC 和正常肺组织的识别能力 目的是识别对肺鳞状细胞癌早期的发生有着至关重要的分子标志物。为进一步揭示 LUSC 发生和发展的机理奠定基础,并对 LUSC 的早期诊断提供理论支持,因此识别的基因必须具有足够高的癌症分类能力,采用的分类模型必须具有高的分类精度。对本文方法筛选得到的分子标志物分别采用支持向量机(SVM)、随机森林(RF)、人工神经网络(ANN)三种分类模型进行建模,分类结果如表 2 所示。

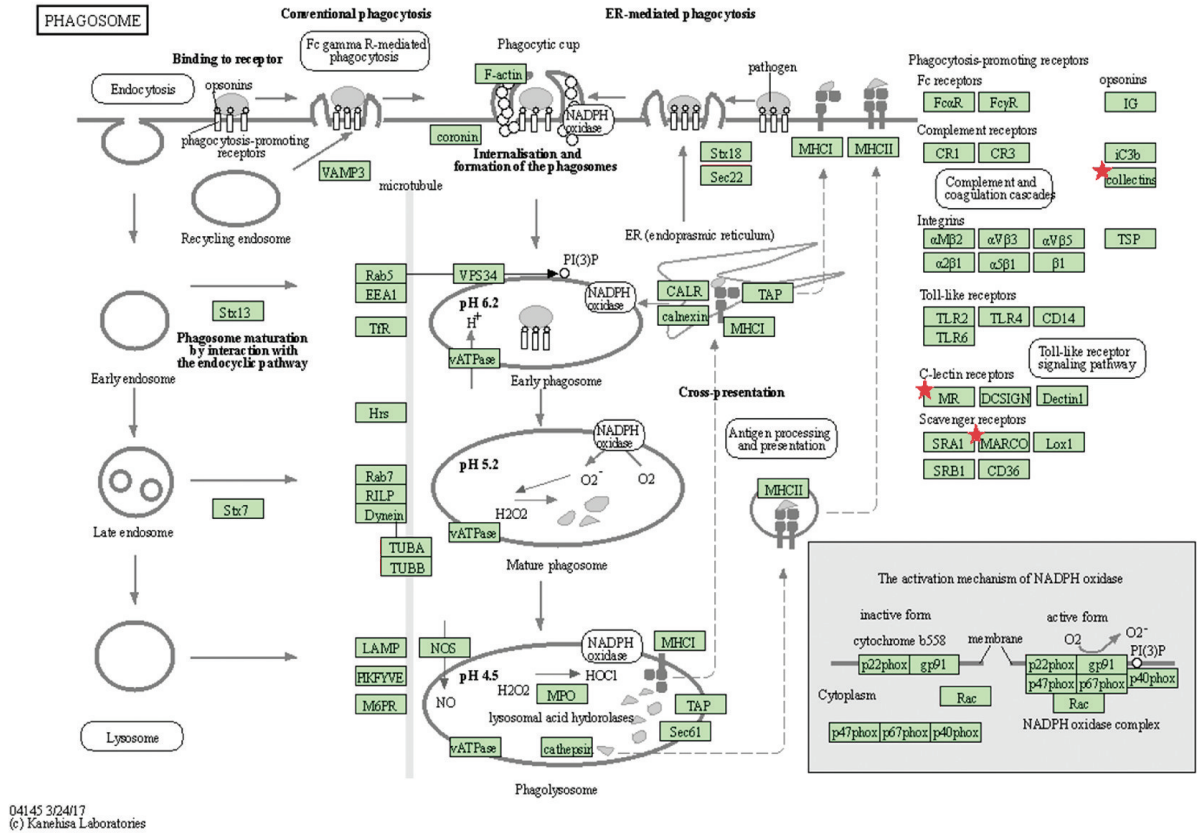

图 6 *MRC*1、*MARCO*、*SFTPA*2 和 *SFTPA*1 在吞噬通路中的位置

Fig.6 Positions of *MRC*1, *MARCO*, *SFTPA*2, and *SFTPA*1 in phagosome

表 2 肺鳞状细胞癌分子标志物的模式识别结果

Table 2 Pattern recognition results of LUSC molecular markers

分子标志物	SVM	RF	ANN
ACC	0. 993 2	0.989 8	0.993 2
SEN	1.000 0	0.995 9	1.000 0
SPE	0. 960 8	0.960 0	0.960 8
MCC	0. 976 2	0.963 6	0.976 2

从表中可以看出,对于三种分类模型分类结果,模型 SVM 和模型 ANN 的敏感性达到 1,说明模型对于癌症 I 期的识别准确率最高;模型 RF 的敏感性达到了 0.99,说明其对癌症 I 期也具有很高的识别准确率。对三种分类模型的结果综合对比可以看出,三种模型的准确率都达到了 99%左右,马修斯相关

系数的值也都达到了 0.96 以上,表明筛选出的分子标志物对肺鳞状细胞癌癌旁和 I 期的分类都具有很高的区分能力,也说明了通过通路富集和基因的生物学功能分析筛选分子标志物的方法具有有效性和可靠性。

(2)对 LUSC 和其余癌症的识别能力 为了验证 LUSC 分子标志物的特异性,即区分 LUSC 和其他癌症的能力,使用支持向量机建立肿瘤分类模型。表 3 显示了 12 个 LUSC 的分子标志物对于 LUSC 样本和甲状腺癌样本、肝癌样本、肾透明细胞癌样本的模式识别结果。

表 3 肺鳞状细胞癌的分子标志物区别肺鳞状细胞癌与其他癌症的结果

Table 3　Results of LUSC molecular markers distinguishing between LUSC and other cancers

Samples	ACC	SEN	SPE	MCC
LUSC-THCA	0.939 7	0.967 3	0.916 1	0.880 9
LUSC-LIHC	0.880 4	0.983 7	0.734 1	0.762 1
LUSC-KIRC	0.880 1	0.979 6	0.790 4	0.777 7

由表 3 可知,LUSC 分子标志物对 LUSC 的识别度很高,敏感性达到 96%以上,在排除甲状腺癌时达到 91.61%的特异性,排除肝癌和肾透明细胞癌的特异性也达到 73%以上,表明使用 12 个 LUSC 分子标志物既能够对 LUSC 样本有极高的识别度,同时也能排除其他癌症样本,验证了本文的筛选方法得到的 LUSC 分子标志物具有特异性。

3.2　Fisher 判别建立模型

为了建立肺鳞状细胞癌早期的分类预测最优模型,为临床早期预测及诊断提供一种新的辅助方法,我们需要对筛选出的分子标志物进行多种组合的 Fisher 判别分析建模,建立模型的基因尽可能少,并且模型必须具有足够高的癌症分类精度。

3.2.1. 单一标志物的判别模型分类结果

对候选基因集中的每个基因都进行建模预测,分类结果见表 4。

表 4　单一标志物的判别模型分类结果

Table 4　Classification results of discriminant model for single marker

Genes	ACC	SEN	SPE	MCC
*SFTPA*1	0.989 8	0.995 9	0.959 2	0.963 0
ESAM	0.983 0	0.987 8	0.959 2	0.939 3
*CLDN*18	0.979 6	0.995 9	0.898 0	0.925 2
*CDH*5	0.976 2	0.991 8	0.898 0	0.912 8
*JAM*2	0.976 2	0.991 8	0.898 0	0.912 8
*SFTPA*2	0.976 2	0.991 8	0.898 0	0.912 8
*F*11	0.969 4	0.991 8	0.857 1	0.886 9
*MRC*1	0.959 2	0.991 8	0.795 9	0.847 5
MARCO	0.955 8	0.983 7	0.816 3	0.835 7
*CD*34	0.952 4	0.987 8	0.775 5	0.821 1
*F*8	0.952 4	0.987 8	0.775 5	0.821 1
CFD	0.932 0	0.959 2	0.795 9	0.755 1

由表 4 中可以看出,在所有的模型中,敏感性都在 0.98 左右,说明其对于癌症 I 期的样本具有较好的识别能力;特异性在 0.77 以上,说明其对癌旁样本的识别能力有些不足。综合来看,建立的 12 个模型的准确率都在 93%以上,其中基因 *SFTPA*1 和基因 *ESAM* 建立的模型准确率最高,都在 98%以上,说明其模型对肺鳞状细胞癌癌旁和癌症 I 期具有很高的区分能力。

3.2.2　两个特征基因组合判别模型的分类结果

对 12 个特征基因进行两两配对建立模型,利用 Fisher 判别分析共建立了 66 个判别模型,取准确率及马修斯相关系数排名前十的判别模型,其预测结果见表 5。

由表 5 可以明显看出,与单基因建立的模型相比,两个特征基因组合的模型预测准确率更高,其诊断价值也更高。在这十个模型中,有 7 个模型的准确率达到了 0.99。其中,前六个模型的特异性,即识别正常肺组织样本的能力达到了 1,而敏感性(识别 LUSC 癌症 I 期样本)也达到了 0.99,说明模型对肺鳞状细胞癌癌旁和癌症 I 期样本都具有很高的识别能力。

表 5 两个特征基因组合模型的分类结果

Table 5 Classification results of discriminant model for two characteristic genes

Genes	ACC	SEN	SPE	MCC
*SFTPA*1、*ESAM*	0.993 2	0.991 8	1.000 0	0.976 2
*SFTPA*1、*CDH*5	0.993 2	0.991 8	1.000 0	0.976 2
ESAM、*SFTPA*2	0.993 2	0.991 8	1.000 0	0.976 2
ESAM、*F*11	0.993 2	0.991 8	1.000 0	0.976 2
ESAM、*MRC*1	0.993 2	0.991 8	1.000 0	0.976 2
*CDH*5、*SFTPA*2	0.993 2	0.991 8	1.000 0	0.976 2
*CLDN*18、*CFD*	0.993 2	0.995 9	0.979 6	0.975 5
*SFTPA*1、*JAM*2	0.989 8	0.991 8	0.979 6	0.963 0
*SFTPA*1、*CD*34	0.989 8	0.991 8	0.979 6	0.963 0
ESAM、*JAM*2	0.989 8	0.991 8	0.979 6	0.963 0

3.2.3 多个基因组合模型的诊断价值评估

将 12 个基因按照准确率的大小,从大到小排列,结果为 *SFTPA*1、*ESAM*、*CLDN*18、*CDH*5、*JAM*2、*SFTPA*2、*F*11、*MRC*1、*MARCO*、*CD*34、*F*8、*CFD*。采用 Fisher 判别依次累加基因建立模型,模型预测准确率如图 7 所示。

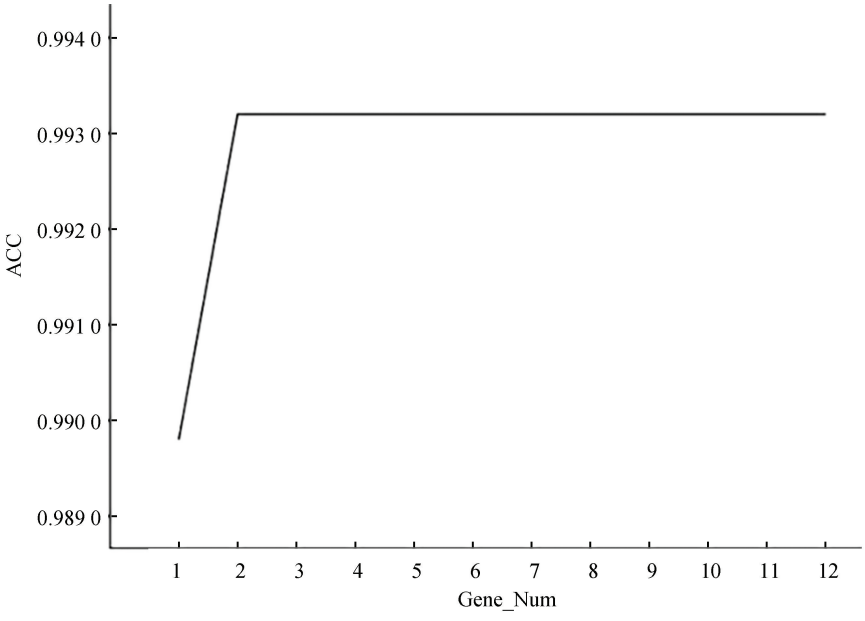

图 7 ACC 与特征基因集大小的关系

Fig.7 Relationship between ACC and size of characteristic genes set

由图 7 可以看出,当基因数量增加为 2 个时,LUSC 的判别模型的分类准确率达到最高,此后当基因再次增加时,其模型的分类预测准确率不在有变化,一直保持稳定。

3.2.4 最优模型及其独立数据集验证

Fisher 判别模型由判别函数来表示,通常采用重心距离来定义类与类之间的距离,从而对待判样本进行判别。通过对上述涉及的多个模型的分类结果对比,最终选择了基因 *SFTPA*1 和 *ESAM*,以及以这两个基因建立的三个分类准确率较高的 LUSC 判别模型,分别是 *SFTPA*1 单基因模型、*ESAM* 单基因模型以及 *SFTPA*1 和 *ESAM* 双基因模型。表 6 显示了三个模型对正常肺组织和 LUSC 癌症 I 期样本的判别重心以及模型的判别函数。综合考虑模型对 LUSC 的分类敏感性和特异性,最终保留 *SFTPA*1 和 *ESAM* 建立的 Fisher 判别模型。

为了验证判别模型的有效性,需要在不同的平台上下载 LUSC 的数据集作为验证集,验证模型对于 LUSC 的癌症分类能力。从 GEO 数据库中下载的 GSE11969 数据集,经过预处理之后,作为独立验证集参与模型的检验。*SFTPA*1 和 *ESAM* 双基因判别模型对独立验证集的分类结果如表 7 所示。

表 6　判别模型

Table 6　Discriminating model

Genes	Centroid of discriminant		Discriminant function
	normal	cancer in stage I	
*SFTPA*1	0.114 1	0.006 9	0.222 6×SFTPA1
ESAM	0.138 7	0.018 2	
*SFTPA*1、*ESAM*	0.229 1	0.022 9	0.247 7×ESA*M*

表 7　独立验证集的分类结果

Table 7 Classification result of independent verification set

Data set	ACC	SEN	SPE	MCC
GSE11969	0.909 1	0.882 4	1.000 0	0.793 9

独立验证集的分类结果显示,LUSC 的判别模型的分类特异性为 1,敏感性为 0.88,说明该模型对 LUSC 癌症样本的识别能力相对较弱,但是对正常样本的识别能力非常好。总体准确率在 90%以上,说明该模型对 LUSC 有较高的分类能力,且普适性较好。

4　结　论

通过基因组百科全书(KEGG)通路富集分析和基因生物学功能分析,结合统计学的相关方法筛选出 12 个 LUSC 发生的早期标志物,对这些基因采用机器学习的方法建模,得到准确率在 98%以上,说明该方法筛选得到的分子标志物对 LUSC 早期癌症样本和正常组织样本的分类能力很好,且筛选得到的 12 个分子标志物(*CLDN*18, *CD*34, *ESAM*, *JAM*2, *CDH*5, *F*11, *F*8, *CFD*, *MRC*1, *MARCO*, *SFTPA*2, *SFTPA*1)可能成为 LUSC 的诊断及治疗靶点,有助于 LUSC 分子机制的研究。其次,本文建立了一种基于早期标志基因的肿瘤预测模型,对于不同平台的 LUSC 数据集,其分类准确率高于 90%,说明该模型的普适性较好,具有良好的临床应用前景。虽然本文的研究具有良好的结果,但是研究内容还是仅仅停留在生物信息学预测层面,缺少实验验证和分析,后续应与临床试验合作开展检验工作,使得结论更加严谨。

致谢:本研究承蒙国家自然科学基金(No. 11572014),国家科技部的重点研发项目(No. 2017YFC0111104)以及智能化生理测量与临床转化北京国际科技合作基地的资助。

参考文献(References)

[1]HAMMERMAN P,LAWRENCE M,VOET D,et al. Comprehensive genomic characterization of squamous cell lung cancers[J]. Nature, 2012, 489(7417): 519-525. DOI: 10.1038/nature11404.

[2]LIU Lei, ZHAO Enhong, LI Chunhui, et al. TRIM28, a new molecular marker predicting metastasis and survival in early-stage non-small cell lung cancer[J]. Cancer Epidemiology,2013,37(1):71-78. DOI: 10.1016/j.canep.2012.08.005.

[3]TSENG R C, LEE S H, HSU H S, et al. SLIT2 attenuation during lung cancer progression deregulates β-catenin and E-cadherin and associates with poor prognosis[J]. Cancer research,2010, 70(2):543-551. DOI: 10.1158/0008-5472.CAN-09-2084.

[4]YU Guiping, CHEN Guoqiang, WU Song, et al. The expression of PEBP4 protein in lung squamous cell carcinoma[J]. Tumour Biology,2011,32(6): 1257-1263. DOI: 10.1007/s13277-011-0230-1.

[5]李建更, 李萍, 严志, 等. 基于机器学习方法的胃癌分型标志基因提取[J]. 中国生物医学工程学报,2009,028(004): 554-560. DOI: 10.3969/j.issn.0258-8021.2009.04.013.

LI Jiangeng, LI Ping, YAN Zhi, et al. Selection of gastric cancer subgroups marker genes based on machine learning methods[J]. Chinese Journal of Biomedical Engineering, 2009,028(004): 554-560. DOI: 10.3969/j.issn.0258-8021.2009.04.013.

[6]ZHANG Fei, WANG Shixiang, WANG Ling, et al. Pattern recognition of the lung squamous cell carcinoma tumor pro-

gression classification model and signature genes identification[J]. Progress in Biochemistry and Biophysics, 2016, 43(1): 63-74. DOI: 10.16476/j.pibb.2015.0352.

[7] FENG Yumei, LI Xiaoqing, SUN Baocun, et al. Evidence for a transcriptional signature of breast cancer[J]. Breast Cancer Research and Treatment, 2010, 122(1): 65-75. DOI: 10.1007/s10549-009-0505-z.

[8] LAU S K, BOUTROS P C, PINTILIE M, et al. Three-gene prognostic classifier for early-stage non-small-cell lung cancer[J]. Journal of Clinical Oncology, 2007, 25(35): 5562-5569. DOI: 10.1200/JCO.2007.12.0352.

[9] WEN Jianxin, WANG Xuedong, LI Xiaoqin, et al. Signature genes identification of the breast cancer occurrence and pattern recognition[J]. Progress in Biochemistry and Biophysics, 2017, 44(11): 1016-1025. DOI: 10.16476/j.pibb.2017.0221.

[10] WANG Xuedong, SHANG Wenhui, CHANG Yu, et al. Methylation signature genes identification of the lung squamous cell carcinoma occurrence and recognition research[J]. Journal of Computational Biology, 2018, 25(10): 1161-1169. DOI: 10.1089/cmb.2018.0069.

[11] KOVAL M. Claudins—key pieces in the tight junction puzzle[J]. Cell Communication & Adhesion, 2006, 13(3): 127-138. DOI: 10.1080/15419060600726209.

[12] MARTIN T A. Tight junctions in cancer metastasis[J]. Seminars in Cell and Developmental Biology, 2014, 36(1): 898-936. DOI: 10.1016/j.semcdb.2014.09.008.

[13] SANADA Y, OUE N, MITANI Y, et al. Down-regulation of the claudin-18 gene, identified through serial analysis of gene expression data analysis, in gastric cancer with an intestinal phenotype[J]. The Journal of Pathology, 2006, 208(5): 633-642. DOI: 10.1002/path.1922.

[14] 王行富, 张声, 郑珂, 等. CLDN18 表达在胃癌进展中的意义[J]. 中国肿瘤临床, 2011, 38(11): 642-646. DOI: 10.3969/j.issn.1000-8179.2011.11.011.
WANG Xingfu, ZHANG Sheng, ZHENG Ke, et al. Significance of CLDN18 expression in gastric cancer progression[J]. Chinese Journal of Clinical Oncology, 2011, 38(11): 642-646. DOI: 10.3969/j.issn.1000-8179.2011.11.011.

[15] 李艳宏, 赵秀芳. 乳腺肿瘤组织间质中 CD34、α-SMA 的表达及意义[J]. 海南医学院学报, 2014, 20(8): 1035-1037. DOI: 10.13210/j.cnki.jhmu.20140416.018.
LI Yanhong, ZHAO Xiufang. Expression of CD34 and α-SMA in breast tumor tissue mesenchyme[J]. Journal of Hainan Medical College, 2014, 20(8): 1035-1037. DOI: 10.13210/j.cnki.jhmu.20140416.018.

[16] 梁劲松, 宋文林, 吴艳. 内皮细胞选择性黏附分子在糖尿病肾病中的作用机制[J]. 中国糖尿病杂志, 2016, 24(4): 362-366. DOI: 10.3969/j.issn.1006-6187.2016.04.018.
LIANG Jinsong, SONG Wenlin, WU Yan. The study on the molecular mechanism of endothelial cells in diabetic nephropathy[J]. Chinese Journal of Diabetes, 2016, 24(4): 362-366. DOI: 10.3969/j.issn.1006-6187.2016.04.018.

[17] KOK-SIN T, MOKHTAR N M, ALI HASSAN N, et al. Identification of diagnostic markers in colorectal cancer via integrative epigenomics and genomics data[J]. Oncology Reports, 2015, 34(1): 22-32. DOI: 10.3892/or.2015.3993.

[18] LABELLE M, SCHNITTLER H J, AUST D E, et al. Vascular endothelial cadherin promotes breast cancer progression via transforming growth factor β signaling[J]. Cancer Research, 2008, 68(5): 1388-1397. DOI: 10.1158/0008-5472.CAN-07-2706.

[19] JOHANN D J, WEI B R, PRIETO D R A, et al. Combined blood/tissue analysis for cancer biomarker discovery: Application to renal cell carcinoma[J]. Analytical Chemistry, 2010, 82(5): 1584-1588. DOI: 10.1021/ac902204k.

[20] HENDRIX M J, SEFTOR E A, MELTZER P S, et al. Expression and functional significance of VE-cadherin in aggressive human melanoma cells: Role in vasculogenic mimicry[J]. Proceedings of the National Academy of Sciences of the United States of America, 2001, 98(14): 8018-8023. DOI: 10.1073/pnas.131209798.

[21] ZHANG Lizhi, MEI Jiong, QIAN Zhikang, et al. The role of VE-cadherin in osteosarcoma cells[J]. Pathology Oncology Research, 2010, 16(1): 111-117. DOI: 10.1007/s12253-009-9198-1.

[22] LIMA L G, MONTEIRO R Q. Activation of blood coagulation in cancer: Implications for tumour progression[J]. Bioscience Reports, 2013, 33(5): e00064. DOI: 10.1042/BSR20130057.

[23] 刘空前. 肝病患者血清补体 C3 检测的结果分析和意义[J]. 实验与检验医学, 2013, (6): 633-634. DOI: 10.3969/j.issn.1674-1129.2013.06.055.
LIU Kongqian. Analysis of the results and significance of serum complement C3 in patients with liver disease[J]. Experimental and Laboratory Medicine, 2013, (6): 633-634. DOI: 10.3969/j.issn.1674-1129.2013.06.055.

[24] MANTOVANI A, ROMERO P, PALUCKA A K, et al. Tumour immunity: Effector response to tumour and role of the microenvironment[J]. Lancet, 2008, 371(9614): 771-783. DOI: 10.1016/S0140-6736(08)60241-X.

[25] 章必成, 赵勇, 王俊, 等. 人肺腺癌肿瘤相关巨噬细胞的活化表型鉴定[J]. 武汉大学学报: 医学版, 2009, 30(05): 595-598+602+706. DOI: 10.14188/j.1671-8852.

2009.05.025.
ZHANG Bicheng, ZHAO Yong, WANG Jun, et al. Activated phenotypes identification of tumor-associated macrophages in human lung adenocarcinoma[J]. Medical Journal of Wuhan University, 2009, 30(05): 595 - 598 + 602 + 706. DOI: 10.14188/j.1671-8852.2009.05.025.

[26]NEWMAN S L, HOLLY A. Candida albicans is phagocytosed, killed, and processed for antigen presentation by human dendritic cells[J]. Infection and Immunity, 2001, 69(11): 6813 - 6822. DOI: 10.1128/IAI.69.11.6813 - 6822.2001.

[27]郝嘉, 肖颖彬. 肺表面活性物质相关蛋白 A 研究现状[J]. 中国危重病急救医学, 2000, 12(1): 60-61. DOI: 10.3760/j.issn:1003-0603.2000.01.028.
HAO Jia, XIAO Yingbin. Research status of pulmonary surfactant associated protein A[J]. Chinese Critical Care Medicine, 2000, 12(1): 60 - 61. DOI: 10.3760/j.issn:1003-0603.2000.01.028.

[28]陈群娥, 徐艳, 刘建平, 等. 肺表面活性物质在预防新生儿呼吸窘迫综合征中的临床价值[J]. 广西医科大学学报, 2016, 33(2): 317-319. DOI: 10.16190/j.cnki.45-1211/r.2016.02.038.
CHEN Qune, XU Yan, LIU Jianping, et al. Clinical value of pulmonary surfactant in the prevention of neonatal respiratory distress syndrome[J]. Journal of Guangxi Medical University, 2016, 33(2): 317-319. DOI: 10.16190/j.cnki.45-1211/r.2016.02.038.

[29]WANG Shanmei, HE Xian, LI Nan, et al. A novel nanobody specific for respiratory surfactant protein A has potential for lung targeting[J]. International Journal of Nanomedicine, 2015, 10: 2857-2869. DOI: 10.2147/IJN.S77268.

[责任编辑:吴永英]