[DOI]10.12016/j.issn.2096-1456.2022.04.004

·基础研究·

基于主成分分析和差异表达基因构建口腔鳞状细胞癌诊断模型

温凌杜[1,2]，　王子弘[3]，　张国明[2]，　赖茜[2]，　杨宏宇[4]
1. 广州医科大学研究生院，广东 广州（510000）；　2. 深圳市宝安中医院（集团）口腔科，广东 深圳（518000）；
3. 深圳市宝安妇幼保健院口腔科，广东 深圳（518000）；　4. 北京大学深圳医院口腔科，广东 深圳（518000）

【摘要】　目的　探讨以主成分分析（principal component analysis，PCA）法分析口腔鳞状细胞癌（oral squamous cell carcinoma，OSCC）的差异表达基因（differentially expressed genes，DEGs）数据库构建的OSCC诊断模型的价值，为临床诊疗提供参考。**方法**　从癌症基因组图谱（The Cancer Genome Atlas，TCGA）数据库中获取OSCC和正常对照样本的RNA-seq表达数据，通过R软件对表达数据进行归一化和差异表达分析，以筛选出DEGs，并同时对DEGs行基因本体（gene ontology，GO）和京都基因与基因组百科全书（Kyoto Encyclopedia of Genes and Genomes，KEGG）富集分析，以发现主要生物学特征。随机选取RNA-seq中DEGs表达数据的70%作为训练集以及30%作为测试集后，应用PCA法对训练集数据进行分析，提取与诊断OSCC相关的主成分（principal components，PC）构建PCA模型，再分别绘制训练集和测试集PCA模型的受试者工作特征（receiver operating characteristic，ROC）曲线并计算曲线下面积（area under curve，AUC），以评估PCA模型对OSCC诊断的准确性。**结果**　从TCGA数据库中获取OSCC和正常对照样本的RNA-seq表达数据分别为330例、32例。以错误发现率（false discovery rate，FDR）＜0.001和|log2FC|（|log2 fold change|）＞4为阈值，共筛选出159个下调和248个上调DEGs，主要富集在中间纤维、黑素体膜等细胞成分，以及色素和唾液相关的生物过程；主要参与唾液分泌、酪氨酸代谢等通路（*P*.adjust＜0.05和*Q*＜0.05）。将DEGs拟作为诊断OSCC的肿瘤标志物，对训练集行PCA分析显示，主成分前3位PC1、PC2、PC3方差的贡献率分别为0.873、0.100、0.023，三者累计方差的贡献率为0.996，主成分前3位PC1、PC2、PC3包含颌下腺雄激素调节蛋白3B（submaxillary gland androgen regulated protein 3B，SMR3B）、富含脯氨酸27（proline rich 27，PRR27）、组蛋白3（histatin 3，HTN3）、抗凝素（statherin，STATH）、胱抑素D（cystatin D，CST5）、包含A族成员2的BPI折叠（BPI fold containing family A member 2，BPIFA2）、富含脯氨酸的蛋白质Hae Ⅲ亚家族2（proline rich protein Hae Ⅲ subfamily 2，PRH2）、角蛋白35（keratin 35，KRT35）、组蛋白1（histatin 1，HTN1）、淀粉酶α1B（amylase alpha 1B，AMY1B）。进一步结合三者的特征向量构建OSCC的PCA诊断模型，在训练集和测试集ROC曲线中显示该模型的AUC值分别为0.852、0.844，均高于其他基因。**结论**　基于PCA法和DEGs构建的以SMR3B、PRR27、HTN3、STATH、CST5、BPIFA2、PRH2、KRT35、HTN1和AMY1B表达水平为基础的OSCC诊断模型有较高的诊断优势，可为OSCC的早期基因诊断以及PCA模型在临床诊断中的应用提供理论依据。

【关键词】　口腔鳞状细胞癌；　差异表达基因；　肿瘤标志物；　早期诊断；　基因诊断；　主成分分析；　诊断模型；　生物信息学
【中图分类号】　R78　**【文献标志码】**　A　**【文章编号】**　2096-1456（2022）04-0251-07
【引用著录格式】　温凌杜，王子弘，张国明，等. 基于主成分分析和差异表达基因构建口腔鳞状细胞癌诊断模型[J]. 口腔疾病防治，2022，30(4)：251-257. doi：10.12016/j.issn.2096-1456.2022.04.004.

微信公众号

Construction of a diagnostic model for oral squamous cell carcinoma based on principal component analysis and differentially expressed genes　WEN Lingdu[1,2], WANG Zihong[3], ZHANG Guoming[2], LAI Xi[2], YANG Hon-

【收稿日期】2021-06-19；　**【修回日期】**2021-12-28
【基金项目】广东省自然科学基金项目（2019A1515011911）；广东省高水平临床重点专科项目（SZGSP008）；深圳市医疗卫生三名工程（SZSM201512036）
【作者简介】温凌杜，主治医师，硕士，Email：454303328@qq.com
【通信作者】杨宏宇，主任医师，博士，Email：tldxdgl@hotmail.com，Tel：86-755-83923333

gyu[4].　1. Graduate School of Guangzhou Medical University, Guangzhou 510000, China;　2. Department of Stomatology, Shenzhen Baoan Hospital of Traditional Chinese Medicine (Group), Shenzhen 518000, China;　3. Department of Stomatology, Shenzhen Baoan Maternity and Child Health Hospital, Shenzhen 518000, China;　4. Department of Stomatology, Peking University Shenzhen Hospital, Shenzhen 518000, China
Corresponding author: YANG Hongyu, Email: tldxdgl@hotmail.com, Tel: 86-755-83923333

【Abstract】　Objective　To explore the value of an oral squamous cell carcinoma (OSCC) diagnostic model constructed by using principal component analysis (PCA) to analyze a database of differentially expressed genes in OSCC and to provide a reference for clinical diagnosis and treatment. **Methods**　RNA-seq expression data of OSCC and normal control samples were obtained from The Cancer Genome Atlas (TCGA) database, and then, normalized and differentially expressed genes (DEGs) were identified by R software. DEGs were enriched by Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis to identify their main biological characteristics. 70% of DEGs expression data in RNA-seq were randomly selected as the training set and 30% were selected as the test set. Then, the PCA method was applied to analyze the training set data and extract the principal components (PCs) related to the diagnosis of OSCC in order to construct a PCA model. Then, the receiver operating characteristic (ROC) curves of PCA models in the training set and the test set were respectively drawn, and the area under curve (AUC) was calculated to evaluate the accuracy of the PCA model in the diagnosis of OSCC. **Results**　RNA-seq expression data of OSCC and normal control samples obtained from TCGA database included 330 samples and 32 samples, respectively. Using false discovery rate (FDR) <0.001 and |log2 fold change| (|log2FC|) >4 as the thresholds, a total of 159 downregulated and 248 upregulated DEGs were identified, which were mainly enriched in cellular components such as intermediate fiber and melanosomal membrane, pigment and salivation-related biological processes and mainly involved in salivary secretion and tyrosine metabolism pathways (P.adjust<0.05 and Q<0.05). The DEGs were proposed as tumor markers for OSCC, and PCA analysis of the training set showed that the cumulative ratio of variance of PC1, PC2 and PC3: [including submaxillary gland androgen regulated protein 3B (SMR3B), proline rich 27 (PRR27), histatin 3 (HTN3), statherin (STATH), cystatin D (CST5), BPI fold containing family A member 2 (BPIFA2), proline rich protein Hae Ⅲ subfamily 2 (PRH2), keratin 35 (KRT35), histatin 1 (HTN1), amylase alpha 1B (AMY1B)] were 0.873, 0.100 and 0.023, respectively, and the total weight of the three was 0.996. The PCA diagnostic model of OSCC was further constructed by combining the eigenvectors of the above three components. The ROC curves of the training set and test set showed that the AUC values of the PCA model were 0.852 and 0.844, respectively, which were higher than those of other single genes. **Conclusion**　The OSCC diagnostic model based on the expression levels of SMR3B, PRR27, HTN3, STATH, CST5, BPIFA2, PRH2, KRT35, HTN1 and AMY1B constructed with the PCA method and DEGs has a high diagnostic advantage. This study provides a theoretical basis for the early genetic diagnosis of OSCC and the application of the PCA model in clinical diagnosis.

【Key words】　oral squamous cell carcinoma;　differentially expressed genes;　tumor markers;　early diagnosis;　genetic diagnosis;　principal component analysis;　diagnostic model;　bioinformatics
J Prev Treat Stomatol Dis, 2022, 30(4): 251-257.
【Competing interests】　The authors declare no competing interests.
This study was supported by the grants from Natural Science Foundation of Guangdong Province (No. 2019A1515011911); Guangdong Province High-level Clinical Key Specialty (No. SZGSP008); Shenzhen Municipal Medical and Health Project (No. SZSM201512036)

口腔鳞状细胞癌(oral squamous cell carcinoma，OSCC)为头颈部最常见的肿瘤之一，具有较强的侵袭性，常可导致局部浸润以及颈部早期淋巴结转移的发生[1-2]。鉴于OSCC独特的解剖位置以及分子发病机制的多样性，目前临床对于OSCC的治疗通常由外科、肿瘤内科或放疗科等组成的多学科团队来实施个性化的综合治疗方案，尽管在包括放化疗、靶向或手术治疗等方式中取得了较大的进展，但OSCC的发病率和死亡率在过去十年并没有得到显著改善，患者的总体5年生存率依然较低[3]，因此在OSCC早期诊断或筛查方面需要新的肿瘤标志物。主成分分析(principal component analysis，PCA)为一种广泛用于医学领域识别模式的多元统计方法，可对影响特定现象的因素进行分类，或通过切割方差较小的主成分(principal component，PC)以降低维数，从而筛选出可用于开发新模型的PC，而PC的权重则可用于计算每个因素在数据中的贡献[4-5]。本研究拟通过TCGA数据

库筛选出OSCC患者的差异表达基因(differentially expressed genes,DEGs)数据,并应用PCA法来确定可用于OSCC诊断的主要因素并以此构建诊断模型,以期为OSCC的早期基因诊断以及PCA模型在临床诊断中的应用提供理论依据。

1　资料和方法

1.1　微阵列数据收集与处理

从TCGA数据库中,选择HTSeq-FPKM工作流程,并以"other and unspecified parts of tongue、other and unspecified parts of mouth、floor of mouth、gum、lip、palate、base of tongue、other and ill-defined sites in lip、oral cavity and pharynx"为检索条件,获取截止于2021年6月2日数据库中OSCC样本与正常对照样本的RNA-seq表达数据。通过Ensembl数据库提供的"Homo_sapiens.GRCh38.104.chr.gtf.gz"文件行基因名称注释。

1.2　DEGs的筛选

应用limma R软件包对RNA-seq表达数据行归一化处理,并以错误发现率(false discovery rate,FDR)<0.001和|log2FC|>4为具有统计学意义的阈值,行差异基因表达分析筛选出DEGs,结果通过ggplot2 R软件包绘制火山图可视化。

1.3　DEGs的富集分析

应用基因本体论(gene ontology,GO)和京都基因与基因组百科全书(Kyoto Encyclopedia of Genes and Genomes,KEGG)数据库,通过clusterProfiler、org.Hs.eg.db、enrichplot和ggplot2 R软件包,取*P*.adjust<0.05和*Q*<0.05为筛选条件对DEGs行富集分析,以发现DEGs的主要生物学特征并绘制气泡图将结果可视化。

1.4　PCA分析与OSCC诊断模型的构建

将DEGs拟作为诊断OSCC的肿瘤标志物,随机选取RNA-seq中DEGs表达数据的70%作为训练集和30%作为测试集。训练集数据通过prcomp R函数行PCA分析,明确PC的特征向量和权重信息并构建OSCC的诊断模型,其中PC权重碎石图由ggplot2 R软件包绘制。最后通过pROC R软件包分别绘制训练集、测试集PCA模型的受试者工作特征(receiver operating characteristic,ROC)曲线并计算曲线下面积(area under curve,AUC),其中AUC在0.5~0.7时为低准确性,AUC在0.7~0.9时为较高准确性,AUC在0.9以上时为高准确性[6],以评估PCA模型对OSCC的诊断优势。

2　结　果

2.1　微阵列数据的整理与DEGs的筛选

从TCGA数据库中共获取OSCC样本RNA-seq表达数据330例,正常对照RNA-seq表达数据32例。对表达文件的原始微阵列数据进行处理和差异表达分析后,基于FDR<0.001和|log2FC|>4的截止标准总共筛选出组蛋白1(histatin 1,HTN1)、富含脯氨酸27(proline rich 27,PRR27)、组蛋白3(histatin 3,HTN3)、包含A族成员2的BPI折叠(BPI fold containing family A member 2,BPIFA2)、胱抑素D(cystatin D,CST5)和富含脯氨酸的蛋白质HaeIII亚家族2(proline rich protein Hae Ⅲ subfamily 2,PRH2)等159个下调DEGs和MAGE家族成员A10(MAGE family member A10,MAGEA10)、溴域睾丸相关(bromodomain testis associated,BRDT)、G2/M期特异性E3泛素蛋白连接酶(G2/M-phase specific E3 ubiquitin protein ligase,G2E3)、亚精胺/精胺N1-乙酰转移酶像1(spermidine/spermine N1-acetyl transferase like 1,SATL1)、脂肪酰基辅酶A还原酶2假基因1(fatty acyl-CoA reductase 2 pseudogene 1,FAR2P1)和钙结合蛋白1(calbindin 1,CALB1)等248个上调DEGs,数据结果由火山图行可视化处理(图1)。

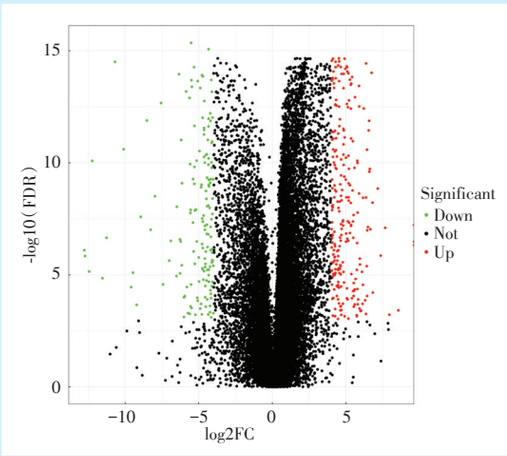
2.2　DEGs的GO与KEGG富集分析

GO注释(图2)显示HTN1、PRR27、HTN3、BPIFA2、CST5和PRH2等407个DEGs主要富集的细胞成分(cellular component,CC)为中间纤维、黑素体膜、几丁质酶体和色素颗粒膜,以及色素和唾液相关的生物过程(biological process,BP),如发育性色素沉着、黑色素生物合成、黑色素代谢和唾液分泌。KEGG通路富集分析(图3)显示DEGs主要参与唾液分泌、酪氨酸代谢、淀粉和蔗糖代谢通路。结果均*P*.adjust<0.05和*Q*<0.05。

2.3　PCA分析与OSCC诊断模型的构建

将DEGs拟作为诊断OSCC的肿瘤标志物,对训练集行PCA分析。由PC所占权重可见(表1、图4),PC1、PC2、PC3方差的贡献率分别为0.873、0.100、0.023,三者累计方差的贡献率为0.996,而随着PC的增多其累计方差的贡献率改变较小,故研究选取主成分前三,即PC1、PC2和PC3用于构建OSCC的诊断模型。

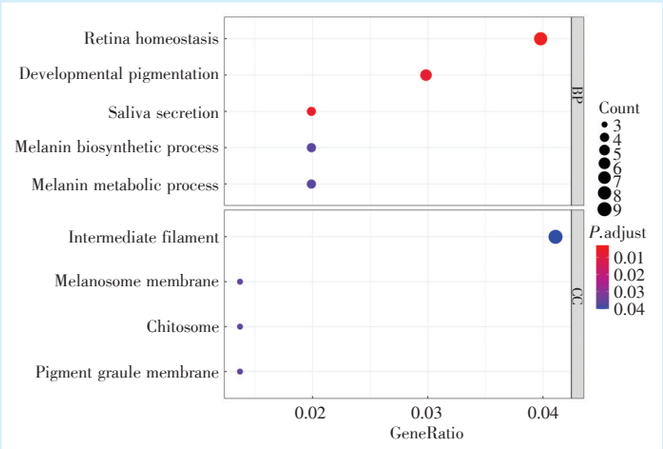
进一步结合PC1、PC2和PC3的特征向量(表2),构建以颌下腺雄激素调节蛋白3B(submaxillary gland androgen regulated protein 3B,SMR3B)、



A total of 159 downregulated DEGs (green dots) such as HTN1, PRR27, HTN3, BPIFA2, CST5 and PRH2 and 248 upregulated DEGs (red dots) such as MAGEA10, BRDT, G2E3, SATL1, FAR2P1 and CALB1 were identified. DEGs: differentially expressed genes; OSCC: oral squamous cell carcinoma; HTN1: histatin 1; PRR27: proline rich 27; HTN3: histatin 3; BPIFA2: BPI fold containing family A member 2; CST5: cystatin D; PRH2: proline rich protein HaeIII subfamily 2; MAGEA10: MAGE family member A10; BRDT: bromodomain testis associated; G2E3: G2/M-phase specific E3 ubiquitin protein ligase; SATL1: spermidine/spermine N1-acetyl transferase like 1; FAR2P1: fatty acyl-CoA reductase 2 pseudogene 1; CALB1: calbindin 1

Figure 1　Volcano map of DEGs distribution in OSCC

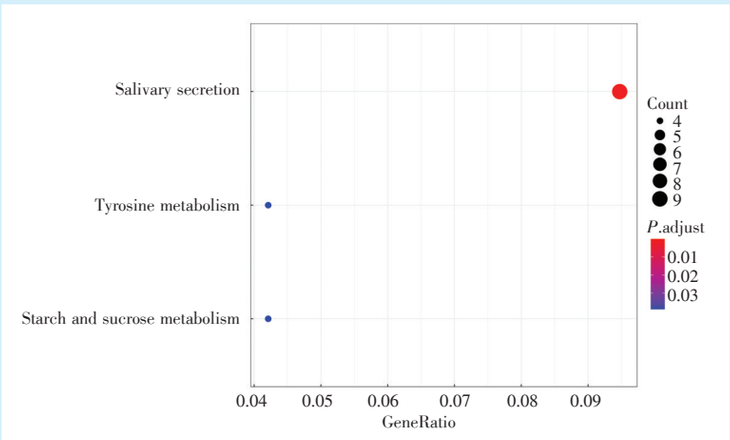
图1　OSCC中DEGs分布的火山图



GO: gene ontology; DEGs: differentially expressed mRNAs; *P*. adjust <0.05 and *Q* < 0.05

Figure 2　GO enrichment analysis of DEGs

图2　DEGs的GO富集分析



KEGG: Kyoto Encyclopedia of Genes and Genomes; DEGs: differentially expressed mRNAs. *P*. adjust < 0.05 and *Q* < 0.05

Figure 3　KEGG pathway enrichment analysis of DEGs

图3　DEGs的KEGG通路富集分析

PRR27、HTN3、抗凝素(statherin，STATH)、CST5、BPIFA2、PRH2、角蛋白35(keratin 35，KRT35)、HTN1和淀粉酶α1B(amylase alpha 1B，AMY1B)表达水平为基础的OSCC诊断模型，模型方程如下。

PC1=SMR3B×(-0.333)+PRR27×(-0.315)+HTN3×(-0.335)+STATH×(-0.338)+CST5×(-0.336)+

BPIFA2×(-0.333)+PRH2×(-0.335)+KRT35×(-0.025)+HTN1×(-0.337)+AMY1B×(-0.337)

PC2=SMR3B×(-0.073)+PRR27×(-0.018)+HTN3×0.031+STATH×0.011+CST5×0.028+BPIFA2×0.031+PRH2×0.029+KRT35×(-0.995)+HTN1×0.023+AMY1B×0.010

PC3=SMR3B×(0.317)+PRR27×0.749+HTN3×(-0.286)+STATH×0.053+CST5×(-0.203)+BPIFA2×(-0.308)+PRH2×(-0.258)+KRT35×(-0.070)+HTN1×(-0.162)+AMY1B×0.145

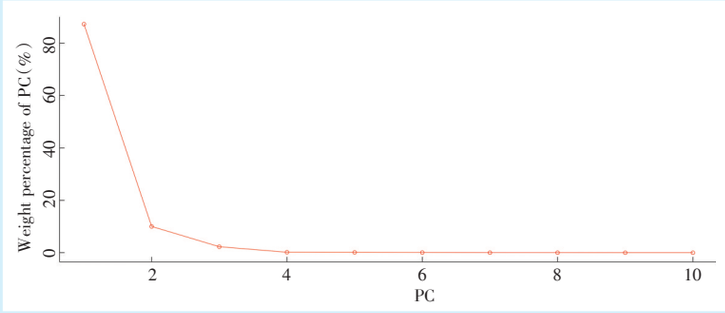
$PC_{综合得分}$=(PC1×0.873+PC2×0.100+PC3×0.023)/0.996

表1　主成分权重信息前五位

Table 1　Weight information about the top five principal components

Project	PC1	PC2	PC3	PC4	PC5
Standard deviation	2.954	1.002	0.479	0.134	0.112
Proportion of variance	0.873	0.100	0.023	0.002	0.001
Cumulative proportion	0.873	0.973	0.996	0.998	0.999

PC: principal components. PC1, PC2 and PC3 includes submaxillary gland androgen regulated protein 3B (SMR3B), proline rich 27 (PRR27), histatin 3 (HTN3), statherin (STATH), cystatin D (CST5), BPI fold containing family A member 2 (BPIFA2), proline rich protein Hae Ⅲ subfamily 2 (PRH2), keratin 35 (KRT35), histatin 1 (HTN1), amylase alpha 1B (AMY1B)



The abscissa represents the principal component, and the ordinate represents the weight percentage of the principal component; PC: principal component

Figure 4　Gravel graph of principal component weight

图4　主成分权重碎石图

表2　PC1、PC2和PC3的特征向量

Table 2　Feature vectors of PC1, PC2 and PC3

Gene	PC1	PC2	PC3
SMR3B	−0.333	−0.073	0.317
PRR27	−0.315	−0.018	0.749
HTN3	−0.335	0.031	−0.286
STATH	−0.338	0.011	0.053
CST5	−0.336	0.028	−0.203
BPIFA2	−0.333	0.031	−0.308
PRH2	−0.335	0.029	−0.258
KRT35	−0.025	−0.995	−0.070
HTN1	−0.337	0.023	−0.162
AMY1B	−0.337	0.010	0.145

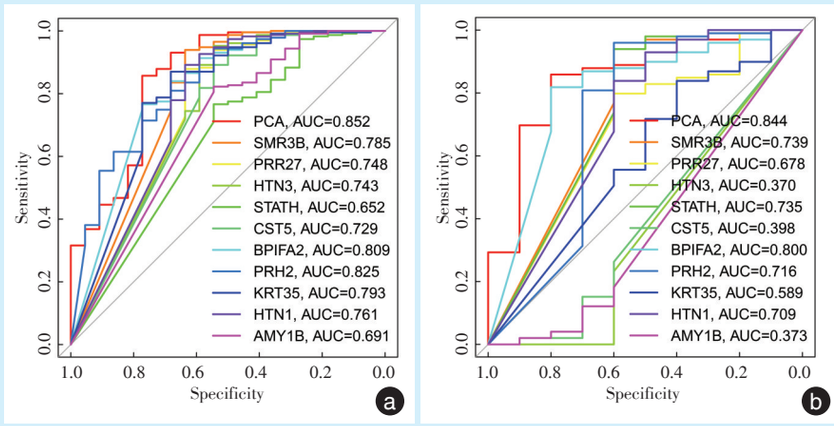
PC: principal components; SMR3B: submaxillary gland androgen regulated protein 3B; PRR27: proline rich 27; HTN3: histatin 3; STATH: statherin; CST5: cystatin D; BPIFA2: BPI fold containing family A member 2; PRH2: proline rich protein HaeIII subfamily 2; KRT35: keratin 35; HTN1: histatin 1; AMY1B: amylase alpha 1B

训练集ROC曲线(图5a)显示PCA模型的AUC值最高(0.852),并且在测试集ROC曲线(图5b)中仍可看到该模型的AUC值(0.844)较其他基因高,表明该模型在OSCC的诊断中具有优势。

3　讨　论

OSCC是最常见的口腔癌类型,具有较高的发病率和恶性程度,可发生在口腔的任何部位,临床以舌前三分之二、上下牙龈以及颊部黏膜较为多见[7]。癌症的筛查或早期诊断被认为是改善预后和提高患者生存率的关键因素[8],口腔独特的解剖位置使临床医师可通过直接目视或触诊等常规检查来评估病变以便对可疑组织进行活检,但作为目前临床常规诊断方法,其对OSCC检测的有效性仍存在争议。研究表明,基于该常规诊断方法仍有大多数OSCC病例在早期阶段未被发现,而该病高死亡率的主要原因是超过50%的OSCC患者首次就诊即被诊断为晚期[9]。在临床工作中也发现,有些患者无法完全张口进行检查,而且OSCC和几种类型口腔潜在恶性疾病具有相似表现,OSCC的诊断在很大程度上依赖于可以识别早期肿瘤变化的临床专业知识,但即便是高年资专业医师也难以完全准确区分口腔潜在恶性疾病和OSCC[10]。因此,需要提高OSCC的早期确诊率以改善患者的治疗效果和预后。

以往研究发现脱落细胞DNA计数、刷拭活检、



a: traning set; b: test set; the PCA model shown in the figure a (AUC value of 0.852) and figure b (AUC value of 0.844) both had the highest value, and the PCA model had a higher diagnostic advantage. PCA: principal component analysis; AUC: area under curve; SMR3B: submaxillary gland androgen regulated protein 3B; PRR27: proline rich 27; HTN3: histatin 3; STATH: statherin; CST5: cystatin D; BPIFA2: BPI fold containing family A member 2; PRH2: proline rich protein HaeIII subfamily 2; KRT35: keratin 35; HTN1: histatin 1; AMY1B: amylase alpha 1B

Figure 5　Training set and test set ROC curves

图5　训练集、测试集ROC曲线

微核分析等技术可用于OSCC早期的诊断，但均存在一定的局限性[11]。肿瘤标志物可用作健康个体和口腔癌临床或组织学阴性患者的筛查工具[12]，并且越来越多的研究表明OSCC涉及多个致癌基因和抑癌基因。活化的蛋白激酶C1受体（receptor for activated C kinase 1，RACK1）通过NF-κB通路增加M2/M1巨噬细胞比率从而促进OSCC进展[13]，KN基序和锚蛋白重复结构域1（KN motif and ankyrin repeat domains 1，Kank1）的异常表达调节Yes相关转录调节蛋白1（Yes1 associated transcriptional regulator，YAP）以促进OSCC中的细胞凋亡并抑制增殖[14]，MiR-92a通过靶向叉头框蛋白P1（forkhead box P1，FOXP1）表达来调控OSCC细胞的生长[15]，这些基因的发现有助于更好地了解OSCC在分子水平的发病机制，也为挖掘可用于早期诊断或筛查OSCC的肿瘤标志物提供了基础。转录组测序（RNA sequencing，RNA-Seq）技术的出现使研究者可以获取OSCC患者的基因表达数据，而当数据集包含大量变量时，PCA作为探索性数据分析的工具，通常用于在构建预测模型之前进行的变量降维，通过数据协方差矩阵的特征值分解或数据矩阵的奇异值分解来执行，可将大量预测的变量减少到几个PC，特别是在嘈杂或具有强相关变量的数据集中[16]。PC则是解释数据方差原始变量的线性组合，线性组合中每个变量对应的系数表示该变量在分量中的相对权重，系数的绝对值越大，对应的变量在计算分量中越重要[17]。Kang等[18]研究发现PCA法可通过对三维计算机断层扫描图像上的大量解剖标志变量分析中识别出最具特征的变量，从而可用于确定哪些解剖结构可最能表征患者的主要变异。秦明丽等[19]基于对132例卵巢癌患者和211例卵巢良性肿瘤患者的血清癌胚抗原（carcinoembryonic antigen，CEA）、糖类抗原125（carbohydrate antigen 125，CA125）、糖类抗原153（carbohydrate antigen 153，CA153）等8项肿瘤标志物建立的PCA-多层感知器（multi perceptronlayer，MPL）-人工神经网络（artificial neural network，ANN）模型研究发现，该模型可有效提升卵巢癌的诊断效能，可为卵巢癌的智能化辅助诊断提供参考。

本研究前期通过TCGA数据库筛选出OSCC与正常对照样本之间的DEGs，拟将其作为OSCC的肿瘤标志物以明确是否可用于其诊断。但即便研究中将DEGs的筛选条件调整为FDR＜0.001和| log2FC |＞4，仍有159个下调DEGs和248个上调DEGs，共407个DEGs被筛选出来。由于这些DEGs中包含数据较大，难以确定哪些基因可作为最能体现诊断OSCC的因素，故应用PCA法对数据进行降维处理。发现PC1、PC2和PC3方差的贡献率分别为0.873、0.100、0.023，三者累计方差的贡献率为0.996。而自PC3后，即便随着PC的增加，累计方差的贡献率值较前三者叠加已变化不大，表明通过PCA法处理后，PC1、PC2、PC3即可代表原DEGs的数据特征，从而可用于OSCC的诊断。研究进一步通过PC1、PC2、PC3累计方差的贡献率和特征向量构建以SMR3B、PRR27、HTN3、STATH、

CST5、BPIFA2、PRH2、KRT35、HTN1 和 AMY1B 表达水平为基础的 OSCC PCA 诊断模型。在训练集和测试集的 ROC 曲线中可以发现，该模型的 AUC 值分别为 0.852、0.844，较模型内其他基因相比表现出明显的诊断优势，并且具有良好的稳定性。尽管该模型在 OSCC 诊断方面显示出其优越性，但对于癌前病变或癌前状态尚未进行具体分析鉴别，且本研究仅基于 TCGA 数据库在生物信息学层面进行，而要应用到中国人群 OSCC 诊断之前，建议结合国内患者人群的数据信息进行验证。

综上所述，本研究基于 PCA 法和 DEGs 构建的以 SMR3B、PRR27、HTN3、STATH、CST5、BPIFA2、PRH2、KRT35、HTN1 和 AMY1B 表达水平为基础的模型对 OSCC 具有较高诊断优势，可为 OSCC 的早期基因诊断以及 PCA 模型在临床诊断中的应用提供理论依据。

【Author contributions】 Wen LD, Wang ZH processed the research, analyzed the data, and wrote the article. Zhang GM, Lai Q assisted the data analysis. Yang HY revised the article and designed the study. All authors read and approved the final manuscript as submitted.

参考文献

[1] Wang Q, Zhi Y, Ren W, et al. Suppression of OSCC malignancy by oral glands derived-PIP identified by iTRAQ combined with 2D LC-MS/MS[J]. J Cell Physiol, 2019, 234(9): 15330-15341. doi: 10.1002/jcp.28180.

[2] 王安训. 表观遗传与口腔鳞状细胞癌[J]. 口腔疾病防治, 2020, 28(10): 613-622. doi: 10.12016/j.issn.2096-1456.2020.10.001.
Wang AX. Epigenetic and oral squamous cell carcinoma[J]. J Prev Treat Stomatol Dis, 2020, 28(10): 613-622. doi: 10.12016/j.issn.2096-1456.2020.10.001.

[3] Ono K, Eguchi T, Sogawa C, et al. HSP-enriched properties of extracellular vesicles involve survival of metastatic oral cancer cells [J]. J Cell Biochem, 2018, 119(9): 7350-7362. doi: 10.1002/jcb.27039.

[4] Giuliani A. The application of principal component analysis to drug discovery and biomedical data[J]. Drug Discov Today, 2017, 22(7): 1069-1076. doi: 10.1016/j.drudis.2017.01.005.

[5] Garcia-Retortillo S, Javierre C, Hristovski R, et al. Principal component analysis as a novel approach for cardiorespiratory exercise testing evaluation[J]. Physiol Meas, 2019, 40(8): 084002. doi: 10.1088/1361-6579/ab2ca0.

[6] Kamarudin AN, Cox T, Kolamunnage-Dona R. Time-dependent ROC curve analysis in medical research: current methods and applications[J]. BMC Med Res Methodol, 2017, 17(1): 53. doi: 10.1186/s12874-017-0332-6.

[7] Shi J, Bao X, Liu Z, et al. Serum miR-626 and miR-5100 are promising prognosis predictors for oral squamous cell carcinoma [J]. Theranostics, 2019, 9(4): 920-931. doi: 10.7150/thno.30339.

[8] Bugshan A, Farooq I. Oral squamous cell carcinoma: metastasis, potentially associated malignant disorders, etiology and recent advancements in diagnosis[J]. F1000Res, 2020, 9: 229. doi: 10.12688/f1000research.22941.1.

[9] Chu HW, Chang KP, Hsu CW, et al. Identification of salivary biomarkers for oral cancer detection with untargeted and targeted quantitative proteomics approaches[J]. Mol Cell Proteomics, 2019, 18(9): 1796-1806. doi: 10.1074/mcp.RA119.001530.

[10] Nagi R, Reddy-Kantharaj YB, Rakesh N, et al. Efficacy of light based detection systems for early detection of oral cancer and oral potentially malignant disorders: systematic review[J]. Med Oral Patol Oral Cir Bucal, 2016, 21(4): e447-e455. doi: 10.4317/medoral.21104.

[11] 刘洋, 高岩, 陈学杰, 等. 脱落细胞DNA定量分析在口腔潜在恶性疾病诊断中的准确性[J]. 北京大学学报(医学版), 2019, 51(1): 16-20. doi: 10.19723/j.issn.1671-167X.2019.01.004.
Liu Y, Gao Y, Chen XJ, et al. DNA cytometry of exfoliated cells in the diagnosis of oral potential malignant disorders[J]. Beijing Da Xue Xue Bao Yi Xue Ban, 2019, 51(1): 16-20. doi: 10.19723/j.issn.1671-167X.2019.01.004.

[12] Lingen MW. Screening for oral premalignancy and cancer: what platform and which biomarkers?[J]. Cancer Prev Res, 2010, 3(9): 1056-1059. doi: 10.1158/1940-6207.CAPR-10-0173.

[13] Dan H, Liu S, Liu J, et al. RACK1 promotes cancer progression by increasing the M2/M1 macrophage ratio *via* the NF-κB pathway in oral squamous cell carcinoma[J]. Mol Oncol, 2020, 14(4): 795-807. doi: 10.1002/1878-0261.12644.

[14] Fan H, Tian H, Cheng X, et al. Aberrant Kank1 expression regulates YAP to promote apoptosis and inhibit proliferation in OSCC [J]. J Cell Physiol, 2020, 235(2): 1850-1865. doi: 10.1002/jcp.29102.

[15] Guo J, Wen N, Yang S, et al. MiR-92a regulates oral squamous cell carcinoma (OSCC) cell growth by targeting FOXP1 expression [J]. Biomed Pharmacother, 2018, 104: 77-86. doi: 10.1016/j.biopha.2018.05.025.

[16] Konishi T, Matsukuma S, Fuji H, et al. Principal component analysis applied directly to sequence matrix[J]. Sci Rep, 2019, 9(1): 19297. doi: 10.1038/s41598-019-55253-0.

[17] Zhang P, West NP, Chen PY, et al. Selection of microbial biomarkers with genetic algorithm and principal component analysis[J]. BMC Bioinformatics, 2019, 20(Suppl6): 413. doi: 10.1186/s12859-019-3001-4.

[18] Kang TJ, Eo SH, Cho H, et al. A sparse principal component analysis of Class III malocclusions[J]. Angle Orthod, 2019, 89(5): 768-774. doi: 10.2319/100518-717.1.

[19] 秦明丽, 王定玉, 王旗, 等. PCA-MPL-ANN模型在卵巢肿瘤良恶性鉴别中的价值[J]. 医学信息, 2021, 34(7): 63-66. doi: 10.3969/j.issn.1006-1959.2021.07.018.
Qin ML, Wang DY, Wang Q, et al. The value of PCA-MPL-ANN model in the differentiation of benign and malignant ovarian tumors[J]. Med Inform, 2021, 34(7): 63-66. doi: 10.3969/j.issn.1006-1959.2021.07.018.

（编辑　张琳，曾曙光）