Biophysics 生物物理学, 2023, 11(1), 1-16
Published Online February 2023 in Hans. https://www.hanspub.org/journal/biphy
https://doi.org/10.12677/biphy.2023.111001

Hans汉斯

肾细胞癌亚型的特征基因筛选及识别研究

付继鹏，韩振波，李晓琴*，王　猛

北京工业大学环境与生命学部，北京

收稿日期：2023年2月28日；录用日期：2023年4月11日；发布日期：2023年4月18日

摘　要

目的：肾细胞癌(Renal Cell Carcinoma, RCC)不是单一疾病，而是几种具有不同遗传驱动因素、临床病程和治疗反应的组织学定义的癌症。为了研究不同肾细胞癌亚型的分子特征，本文提出了一个筛选肾细胞癌亚型分类的流程图，并对亚型的特征基因进行了分析。方法：基于DNA甲基化和基因表达数据，综合统计学方法和机器学习方法，筛选了与肾细胞癌三种亚型相关的特征基因集，并构建了一个肾细胞癌亚型分类模型。结果：本文筛选得到了6个分类特征基因；构建的分类器准确性达到96.6%，精确性和敏感性分别是93.4%、94.7%；独立检验集的准确性为93.1%。对肾细胞癌亚型相关的特征基因集进行富集分析发现：肾透明细胞癌主要与免疫系统的负调节和白细胞增殖与黏附等相关通路有关，肾乳头状细胞癌和肾嫌色性细胞癌都与肾脏和肾脏系统的发育有关。结论：本文通过构建分类模型实现了肾细胞癌亚型的三分类，结果对了解肾细胞癌的亚型形成机制及分类诊断治疗具有一定指导意义。

关键词

肾细胞癌，机器学习，基因表达，DNA甲基化

Screening and Identification of Characteristic Genes of Renal Cell Carcinoma Subtypes

Jipeng Fu, Zhenbo Han, Xiaoqin Li*, Meng Wang

Faculty of Environment and Life, Beijing University of Technology, Beijing

Received: Feb. 28th, 2023; accepted: Apr. 11th, 2023; published: Apr. 18th, 2023

Abstract

Purpose: Renal cell carcinoma (RCC) is not a single disease, but several histologically defined can-

*通讯作者。

文章引用: 付继鹏, 韩振波, 李晓琴, 王猛. 肾细胞癌亚型的特征基因筛选及识别研究[J]. 生物物理学, 2023, 11(1): 1-16. DOI: 10.12677/biphy.2023.111001

cers with distinct genetic drivers, clinical course, and response to therapy. To investigate the molecular characteristics of the different RCC subtypes, a flowchart for the classification of the subtypes of RCC is presented, and the signature genes of the subtypes are analyzed. Methods: Based on DNA methylation and gene expression data, integrated statistical methods and machine learning methods, signature gene sets associated with the three subtypes of renal cell carcinoma were screened, and a model for renal cell carcinoma subtype classification was constructed. Results: In this paper, six classification signature genes were obtained; The accuracy of the constructed classifier reached 96.6%, and the precision and sensitivity were 93.4%, 94.7%; The accuracy of the independent test set was 93.1%. By enrichment analysis of the characteristic gene set related to renal cell carcinoma subtypes, it was found that kidney renal clear cell carcinoma is mainly associated with negative regulation of the immune system and related pathways, such as leukocyte proliferation and adhesion. Both Kidney renal papillary cell carcinoma and Kidney Chromophobe are associated with the development of the kidney and the renal system. Conclusion: In this paper, we achieved the three classification of renal cell carcinoma subtypes by constructing a classification model, and the results will be helpful for understanding the mechanism of renal cell carcinoma subtype formation and for classification, diagnosis, and treatment.

Keywords

Renal Cell Carcinoma, Machine Learning, Gene Expression, DNA Methylation

1. 引言

肾细胞癌作为泌尿系统肿瘤中最常见的三大恶性肿瘤之一，其发病率和死亡率一直位于全球男性恶性肿瘤发病率的前几位。根据世界卫生组织国际癌症研究机构(International Agency for Research on Cancer)最新发布的《2020 年全球最新癌症负担数据》显示：2020 年全球男性新发癌症 1007 万例，其中肾癌新发病人数 27 万例，成为了全球男性第九常见癌症[1]。肾细胞癌是一种在病理和分子组织上高度异质的疾病[2]，包含多种肾细胞癌亚型，每个亚型都有不同的组织学特征、遗传改变、临床表现和治疗反应[3]。由于癌症具有较强的隐蔽性，很多病人发现病情时已经是中晚期，这对于癌症的治疗来说非常困难。随着癌症早期筛查诊断技术的完善和推广，癌症患者能够在早期发现病情并及时给予干预和治疗，普遍延长五年生存时间，整体提高生存率。因此，本研究主要针对肾细胞癌早期病例进行分析，以亚型分类的方式了解肾细胞癌亚型的形成机制，在临床上对肾细胞癌的分类诊断治疗具有重要意义。

肾细胞癌的主要亚型包括肾透明细胞癌(Kidney Renal Clear Cell Carcinoma, KIRC)、肾乳头状细胞癌(Kidney Renal Papillary Cell Carcinoma, KIRP)和肾嫌色性细胞癌(Kidney Chromophobe, KICH)，分别占肾细胞癌总病例的 65%、20%和 5% [4]。除了这三大类外，还有透明细胞乳头状癌、黏液性管状癌和梭形细胞癌等几种更为罕见的癌症[5]。对于肾细胞癌亚型的研究前人做了很多分析：Ricketts 等[6]研究了三个亚型的甲基化模式发现，KIRP 相较其他亚型具有较高的甲基化水平。对染色体拷贝数谱进行分析发现 KIRC 表现出染色体 3p 的显著丢失和 5q 的增加，KICH 主要表现出 1、2、6、10、13 及 17 号染色体显著丢失的模式，且 KIRC 的突变率高于 KICH 的三倍[7] [8]。Davis 等[9]发现线粒体代谢途径在 KICH 中过度激活，而在 KIRC 中受到抑制。有相关研究[6] [10]对 *TP53* 和 *PTEN* 的突变进行分析，发现它们的突变降低了 KIRC 和 KIRP 的生存率。这些研究通过组学之间的对比，发现了不同亚型独特的组学表现形式

及分子特征。

肾细胞癌的亚型分类，Han 等[11]利用基于图像的深度学习框架，使用计算机断层扫描获得的图像来区分肾细胞癌的三种主要亚型，得到了 85%的正确率，88%的特异性。Chaudry 等[12]从完整图像中提取基于纹理、形态和小波的特征，用于 RCC 亚型分类，通过选择感兴趣区域，分类准确性达到了 92%。但伴随着肾细胞癌亚型组学数据的积累和丰富，相比于使用图像数据对肾细胞癌分类研究，建立基于分子特征的分类模型成为可能。此外，利用组学数据研究分类问题能够发现癌症发生的独特机制，进而对癌症的治疗提供帮助。例如 Gui 等[13]基于基因表达谱、体细胞突变和 DNA 甲基化确定了 KIRC 的预后良好的基因并找到了出现敏感反应的药物；Deng 等[14]基于基因表达数据确定了与 KIRP 的免疫相关的潜在治疗靶点。

在这项研究中，利用 DNA 甲基化和 mRNA 表达数据对肾细胞癌亚型早期样本的特征基因进行筛选，寻找到显著分类特征基因，建立分类模型进行分类，并使用独立检验集进行验证；同时对每种亚型的特征基因做 GO (Gene Ontology)分析，发现了每种亚型在通路上的不同之处，这为临床上研究肾细胞癌亚型形成机制以及早期诊断治疗提供了理论基础。

2. 材料与方法

2.1. 数据下载与预处理

2.1.1. 数据下载

癌症基因组图谱[15] (The Cancer Genome Atlas)是世界上最大的基因类工程项目，它包括肾细胞癌在内的 30 多种癌症类型。本研究的肾细胞癌数据取自该项目(https://portal.gdc.cancer.gov/)，主要包括了 KIRC、KIRP 和 KICH 的“unstranded”类型的 mRNA 表达数据和 DNA 甲基化数据(表 1)，删除其中临床分期为 II-IV 期的样本。甲基化阵列平台是 illumina Human Methylation450 BeadChip (GPL13534) [16]，甲基化探针注释从 ENCODE project (http://genome.ucsc.edu/ENCODE/downloads.html)下载[17]。

Table 1. mRNA data (DNA methylation) statistics of renal cell carcinoma subtypes
表 1. 肾细胞癌亚型的 mRNA 数据(DNA 甲基化数据)统计

亚型	癌症 I 期	正常样本
KIRC	272 (161)	72 (24)
KIRP	172 (127)	32 (23)
KICH	20 (20)	24 (25)

利用基因表达数据库(Gene Expression Omnibus, GEO, http://www.ncbi.nlm.nih.gov/geo)检索符合条件的肾细胞癌样本集，形成独立检验集。筛选标准：包含“KIRC”、“KIRP”或“KICH”，并包含研究中筛选的所有分类特征基因。本研究通过检索 GEO 数据库，得到了 GSE76351 作为独立检验集，该检验集的注释平台是 GPL11532，包含了 12 例肾透明细胞癌。

2.1.2. 数据预处理

本研究使用的是 mRNA 表达数据和 DNA 甲基化数据。考虑到影响基因表达的 DNA 甲基化主要发生在启动子区域[18]，选择转录起始位点上游 200 bp (TSS200)的基因组区域的探针，并将该区域所有探针的 β 值的均值作为该基因的甲基化水平。

对任一肾细胞癌亚型样本数据集，由于基因数据差异较大并且维度较高，为了减少后续工作的复杂度，删除在所有样本中表达为“0”的基因，对剩余基因进行归一化处理。归一化的具体操作如公式(1)：

$$X_{\text{new}} = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}} \tag{1}$$

其中 X_i 是各样本在某一基因的表达值，$X_{\min}$ 是某一基因在各样本中表达量的最小值，$X_{\max}$ 是某一基因在各样本中表达量的最大值。

2.2. 肾细胞癌不同亚型分类特征基因的筛选

1) 肾细胞癌亚型差异基因筛选。对特定肾细胞癌亚型的癌症样本和正常样本：计算基因表达数据(和甲基化数据)与样本分类标签之间的 Spearman [19] [20]相关系数，筛选绝对值大于 0.5 的基因，并使用单因素方差分析[21] [22]筛选出样本中的差异显著性基因。在基因表达数据中，符合$|\log_2 FC| > 1$ 的基因作为差异表达基因[23] (Differentially Expressed Gene, DEG)；考虑到甲基化数据对基因表达的影响，将符合$|\log_2 FC| > 0.5$ 的基因作为差异甲基化基因(Differentially Methylation Gene, DMG)。

2) 肾细胞癌亚型特征基因筛选。保留基因表达和 DNA 甲基化共有的癌症样本。针对特定肾细胞癌亚型的癌症样本，由于甲基化数据和基因表达数据之间负相关关系，筛选差异甲基化基因和差异表达基因之间负相关的基因，即高甲基化下调基因和低甲基化上调基因；其次，由于 DNA 甲基化与癌症发生密切相关[24]，因此保留超异常甲基化基因($|\log_2 FC| > 2$)及其对应的表达基因。合并上述两部分得到的基因构成特定肾细胞癌亚型特征基因。

3) 分类特征基因筛选。针对肾细胞癌三种亚型的全部癌症样本及两个组学的亚型特征基因数据，利用方差分析筛选保留三种癌症亚型间的显著差异($P < 0.05$，$FDR < 0.05$)特征，并使用弹性网络[25] [26]对其进行分类贡献值的排名，综合排名及其对分类结果的影响确定分类特征基因。

2.3. 分类器的构建及评估

本研究使用支持向量机[27]模型(Support Vector Machine, SVM)，但面对三分类问题，需要在外层嵌套“One-vs-Rest”(OvR)或“One-vs-One”(OvO)的结构构建三分类器[28]。OvR 是在训练过程中取样本中 N 个类别某一类作为单独一类，将其余 N − 1 类作为另一类，但此方法可能会出现负样本的数量远远大于正样本，从而导致样本量失衡，使训练速度缓慢；OvO 是在训练过程中每两个类别之间构建一个分类器，共需要构建 N(N − 1)/2 个分类器，但是此方法避免了正负样本不平衡的问题。因此本文在构建模型时本研究采用了 OvO 策略。

SVM 是一种解最优分类面问题的方法。通过调用 python 中 sklearn 库完成基于支持向量机函数的分类模型的构建。本研究通过分析和测试，最终确定的 SVM 的主要超参数如下：

1) 核函数：目前 SVM 的核函数主要有线性核函数、多项式核函数、Sigmoid 函数和高斯核函数(radial basis function)，针对研究中样本数据线性不可分问题，本研究使用高斯核函数。

2) 惩罚系数 C 和核函数的系数 gamma：惩罚系数是平衡向量的复杂度和误分类率两者间关系的参数；gamma 是 RBF 核函数中表示单个样本对整个分类平面影响的参数。本研究使用网格调参的方法最终确定“C = 30”，“gamma = 3”。

3) 交叉验证 cv：为了得到稳健可靠的模型，并对模型的泛化误差评估，本研究在构建模型时引入了五折交叉验证(cv = 5)。

本文通过混淆矩阵直观展示分类模型的结果，并利用准确性(Accuracy, ACC)、敏感性(Sensitivity, SEN)、精确性(Precision, PRE)和 F1-score (F1)对分类结果进行评估。计算公式如下：

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

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

$$PRE = \frac{TP}{TP + FP} \tag{4}$$

$$F1 = 2 * \frac{SEN * PRE}{SEN + PRE} \tag{5}$$

其中：TP 表示正确分类的正样本的数量，TN 表示正确分类负样本的数量，FP 表示负样本被判定成正样本的数量，FN 表示正样本被判定为负样本的数量。

为了更好的衡量三分类的情况，本文引入了 kappa 系数。它是一种基于混淆矩阵衡量分类精度的指标。通常情况下 kappa 取值在 0~1 之间，可分为五组一致性级别：0.81~1 几乎完全一致，0.61~0.80 高度一致性，0.41~0.60 中等一致性，0.21~0.40 一般一致性以及 0~0.20 极低一致性。计算公式如下：

$$kappa = \frac{p_o - p_e}{1 - p_e} \tag{6}$$

其中：以三分类为例，样本总个数为 n，假设每一类真实样本个数分别为 a_1，a_2，a_3；预测出来的每一类样本个数分别为 b_1，b_2，b_3。p_o 表示总体分类的正确率；$p_e = \left(a_1 * b_1 + a_2 * b_2 + a_3 * b_3\right) / n^2$。

2.4. 功能富集分析

分别对方法 2.2 中得到的 DEGs 和特定肾细胞癌亚型特征基因进行功能富集分析，以探索与肾细胞癌亚型相关的潜在生物学过程。GO 分析包含了生物过程(Biological Processes, BP)、细胞成分(Cellular Components, CC)和分子功能(Molecular Functions, MF)，本研究主要聚焦于生物过程方面。此外，使用在线工具 STRING (https://string-db.org/)对肾透明细胞癌的富集结果进行了蛋白质－蛋白质相互作用网络分析；通过 Cytoscape 软件中“cytoHubba”筛选得到 10 个 hub 基因，分析 hub 基因与 KIRC 的免疫细胞(主要包括 B 细胞、CD8 + T 细胞、CD4 + T 细胞、巨噬细胞、中性粒细胞和树突状细胞)浸润水平的相关性并对 hub 基因进行了简要分析。

2.5. 本研究的流程图

本研究以 RCC 三种主要亚型为研究对象，结合转录组和 DNA 甲基化数据，实现了对这三种亚型的精准分类，具体筛选流程如图 1 所示。

3. 结果与讨论

3.1. 亚型特征基因筛选

将 mRNA 表达数据和 DNA 甲基化数据归一化，利用 2.2 给出的差异基因筛选方法分别得到了三种亚型的 DEGs 和 DMGs，如图 2 所示，纵坐标轴的阈值选择的是 2 (对应 q value = 0.01)；其中红色表示肿瘤中上调(高甲基化)基因，蓝色表示下调(低甲基化)基因。相比于 KICH，KIRC 和 KIRP 的 DEGs 具有更高的显著性差异；三种亚型的基因表达的差异显著性普遍高于 DNA 甲基化。

利用 2.2 给出的肾细胞癌亚型特征基因筛选方法，获得肾细胞癌亚型特征基因。将同一亚型的 DEGs 和 DMGs 匹配，匹配结果见图 3。对 KIRC 亚型，获得了 271 个在 DNA 甲基化和基因表达数据之间具有负相关关系的匹配基因，此外筛选得到了 31 个超异常甲基化基因及其对应的表达基因，得到基因表达和甲基化两组学总计 302 * 2 个(基因表达和 DNA 甲基化)亚型特征基因。同理，从 KIRP 获得了 92 * 2 个亚型特征基因，包括 82 个匹配基因及 10 超异常甲基化基因；从 KICH 中获得了 113 * 2 个亚型特征基因，包括 98 个匹配基因及 15 个超异常甲基化基因。

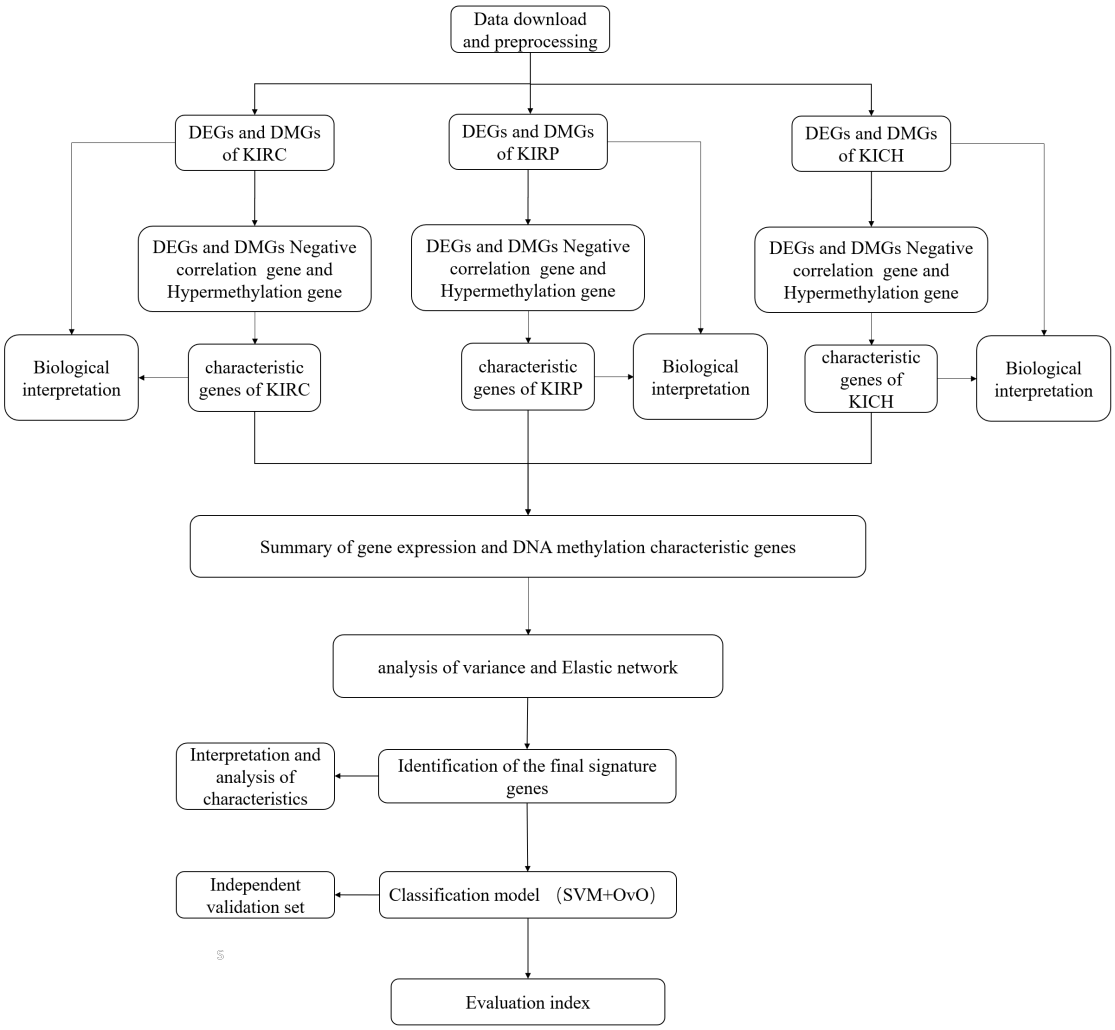


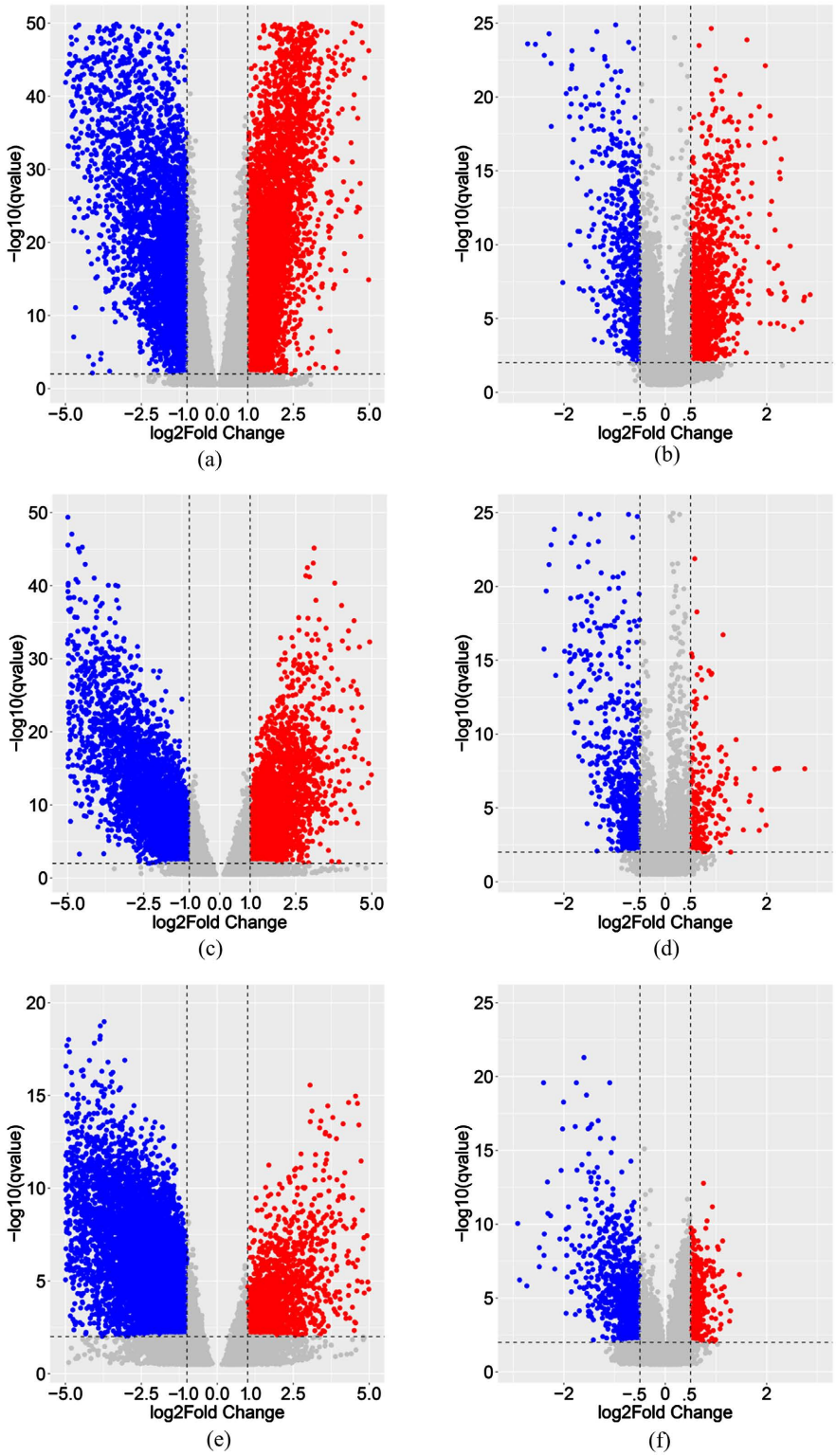
Figure 1. Flowchart of renal cell carcinoma subtype classification
图 1. 肾细胞癌亚型分类流程图

3.2. 分类特征基因筛选

利用 2.2 给出的分类特征基因筛选方法，获得分类特征基因。利用方差分析筛选得到了 852 个显著差异特征基因，利用弹性网络计算显著差异特征基因对分类的贡献，取贡献值排名前 10 的特征基因为候选分类特征基因。如表 2 所示，可以看出 TOP10 的特征基因中只有一个 DNA 甲基化基因(“*M_TMEM88*”)，其余均是 mRNA 表达基因。说明相比于甲基化特征基因，mRNA 表达特征基因对三个亚型分类的贡献度更大。

3.3. 分类模型构建及分类特征基因确定

基于得到的候选分类特征基因，通过构造 SVM 分类模型，实现了对肾细胞癌三种亚型的有监督分类。为了以最少分类特征获得较优分类结果，本研究对不同数目的候选分类特征基因分别进行建模及分类结果比较，结果见表 3 和图 4。结果显示：当分类特征数选择 6 个时，准确性、敏感性和精确性分别为 96.6%、94.7%和 93.4%，F1-score 值和 Kappa 系数达到了 0.94，整体指标均达到了最高。因此，选择前 6 个基因(*PRPS2*, *FGF7*, *TM4SF19*, *PXDNL*, *COQ3*, *SLC25A5*)作为最终的分类特征基因。



(a) and (b)：KIRC 的 DEGs 和 DMGs；(c) and (d)：KIRP 的 DEGs 和 DMGs；(e) and (f)：KICH 的 DEGs 和 DMGs。蓝色表示下调(低甲基化)基因，红色表示上调(高甲基化)基因。

Figure 2. Differentially expressed genes and differentially methylated genes of three subtypes
图 2. 三种亚型的差异表达基因和差异甲基化基因

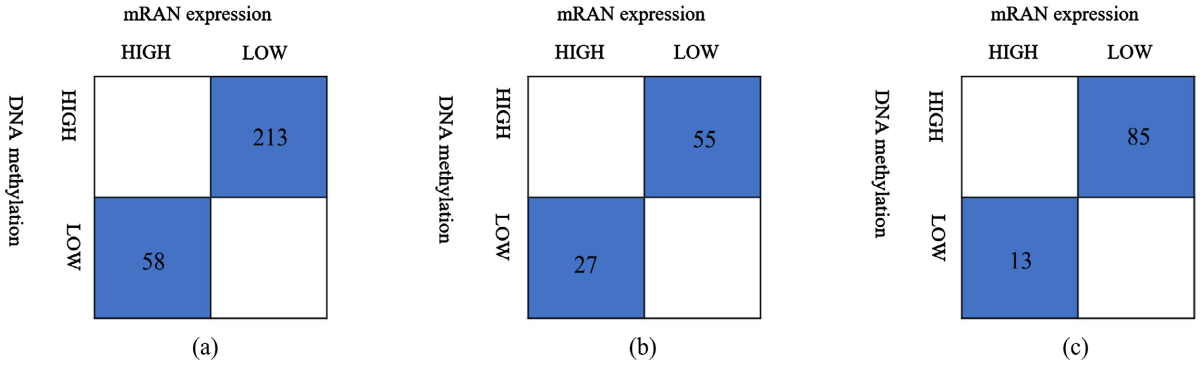


Figure 3. The matching map of gene expression and methylation of Renal Cell Carcinoma. (a) KIRC, (b) KIRP, (c) KICH
图 3. 肾细胞癌基因表达与甲基化的匹配图。(a) KIRC，(b) KIRP，(c) KICH

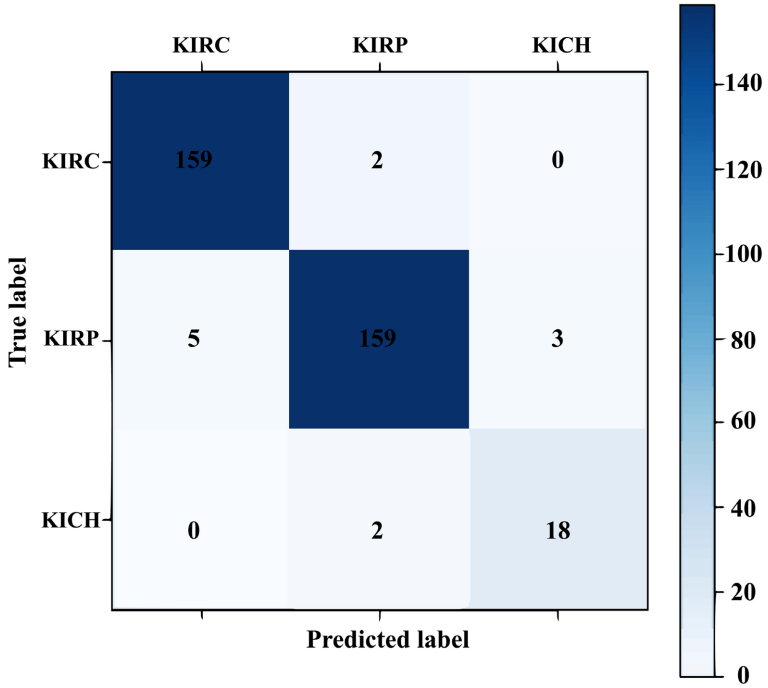


Figure 4. Classification prediction matrix of renal cell carcinoma
图 4. 肾细胞癌分类混淆矩阵

Table 2. Ranking of contribution values of gene
表 2. 基因贡献值排名

TOP	Gene name	TOP	Gene name
1	PRPS2	6	SLC25A5
2	FGF7	7	RHCG
3	TM4SF19	8	PROM2
4	PXDNL	9	M_TMEM88
5	COQ3	10	SCPEP1

* “M_TMEM88”中的“M_”是 DNA 甲基化基因的前缀。

Table 3. Comparison of evaluation indexes of TOP gene classification
表 3. TOP 基因分类评价指标的比较

TOP 基因个数	准确性	精确性	敏感性	F1-score	Kappa 系数
8	96.1%	92.9%	94.2%	0.935	0.92
7	96.5%	93.1%	94.7%	0.939	0.94
6	**96.6%**	**93.4%**	**94.7%**	**0.940**	**0.94**
5	94.0%	88.8%	89.9%	0.893	0.89
4	91.1%	85.0%	82.0%	0.934	0.84

在构建肾细胞癌亚型分类模型时，使用五折交叉验证证明了分类模型的可靠稳定性；为进一步验证亚型分类模型的普适性，对独立检验集 GSE76351 的样本进行识别，模型分类正确率达到 93.1%：即筛选的独立检验集中的 12 个肾透明细胞癌预测正确了 11 个，结果表明分类模型具有一定的普适性，分类特征基因选择比较合理。

分类特征基因在不同亚型中表达量分布见图 5 所示。从图中可以看出：*COQ3* 和 *SLC25A5* 基因在三种亚型中表达的差异较大，*FGF7*、*PXDNL* 和 *TM4SF19* 基因在 KICH 中的表达量较其他两种亚型高，*TM4SF19* 在 KIRC 的表达量低于其他亚型，同时发现 *PXDML* 和 *FGF7* 在 KIRP 表达量最低。其中 *PRPS2*、*COQ3* 和 *SLC25A5* 是 KIRC 的特征基因，且在 KIRC 中都是下调；*FGF7*、*TM4SF19* 和 *PXDNL* 是 KIRP 的特征基因，除了 PXDNL 是上调，其余两个都是下调基因。*FGF7* 是成纤维细胞生长因子(Fibroblast Growth Factor, *FGFs*)家族成员，广泛参与细胞的增殖、分化、迁移、胚胎发育、血管生成以及损伤修复等生物过程[29]。*PRPS2* 是一种嘌呤生物合成途径中的关键限速酶[30]，促进了 MYC 转化细胞中核苷酸生物合成的增加。*PRPS2* 通过 *PRPS2* 内的一个专门的顺式调节元件耦合蛋白质和核苷酸的生物合成 5'UTR，并且它使 Myc 的翻译调节能够直接增加核苷酸生物合成，与癌细胞蛋白质合成率的增加成比例[31]。

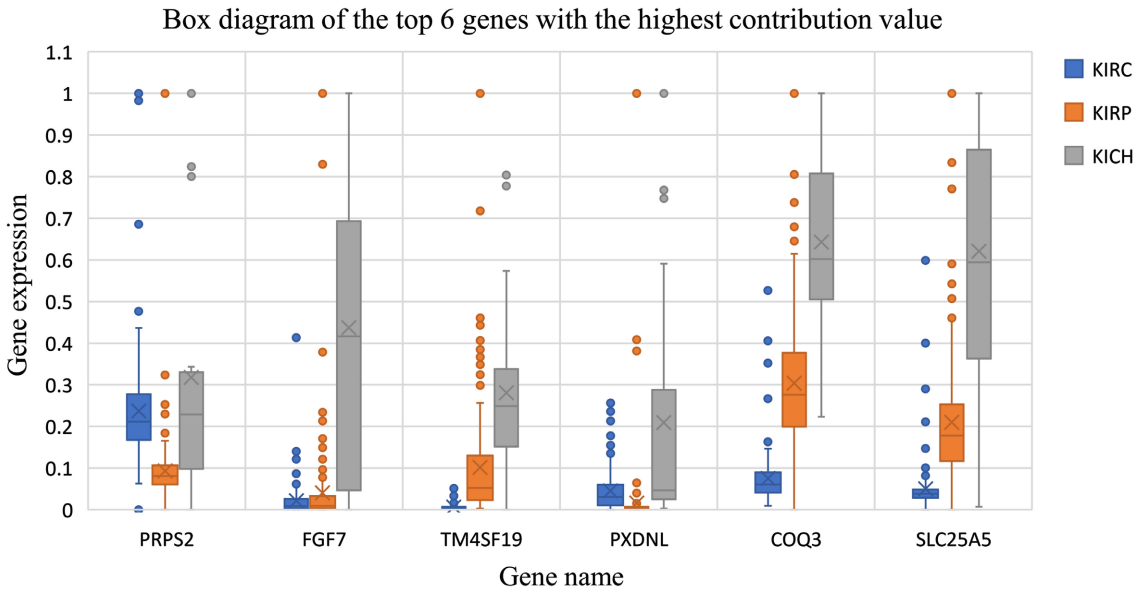


Figure 5. Box diagram of the top 6 genes with the highest contribution value
图 5. 贡献值最高的前 6 个基因的箱式图

3.4. 肾细胞癌亚型基因特征的富集分析

首先，对特定亚型特征基因进行了 GO 分析，表 4 和表 5 分别是 KIRC 和 KIRP 的富集的部分结果。KIRC 富集到的通路主要与钠离子的运输、钠离子运输的正调控、细胞连接组件以及免疫细胞的增殖的相关通路有关；KIRP 主要与突触的组装相关通路有关。对肾细胞癌亚型 DEGs 进行的 GO 分析显示：KIRC 主要与免疫系统的负调节、单核细胞的增殖与分化、白细胞增殖和黏附、淋巴细胞的增殖与分化等相关的免疫系统的相关通路有关(表 S1)；KIRP 主要与肾脏和肾脏系统的发育、钙离子的运输和泌尿生殖系统的发育等相关通路有关(表 S2)；KICH 主要与肾脏和肾脏系统的发育、肾单位的发展以及纤毛组织等相关通路有关(表 S3)。

Table 4. Functional enrichment of KIRC preliminary signature genes
表 4. KIRC 初步特征基因的功能富集

	Description	P value	Count
GO: 0034329	cell junction assembly	2.30E−05	17
GO: 0006814	sodium ion transport	1.73E−06	14
GO: 0035725	sodium ion transmembrane transport	9.41E−06	11
GO: 0050670	regulation of lymphocyte proliferation	9.95E−05	11
GO: 0032944	regulation of mononuclear cell proliferation	0.000108	11
GO: 0070663	regulation of leukocyte proliferation	0.000213	11
GO: 0007043	cell-cell junction assembly	9.02E−05	9
GO: 0070830	bicellular tight junction assembly	5.46E−05	7
GO: 0120192	tight junction assembly	6.42E−05	7
GO: 0120193	tight junction organization	8.10E−05	7
GO: 0002028	regulation of sodium ion transport	0.000101	7
GO: 0043297	apical junction assembly	0.000101	7
GO: 0010765	positive regulation of sodium ion transport	6.61E−05	5

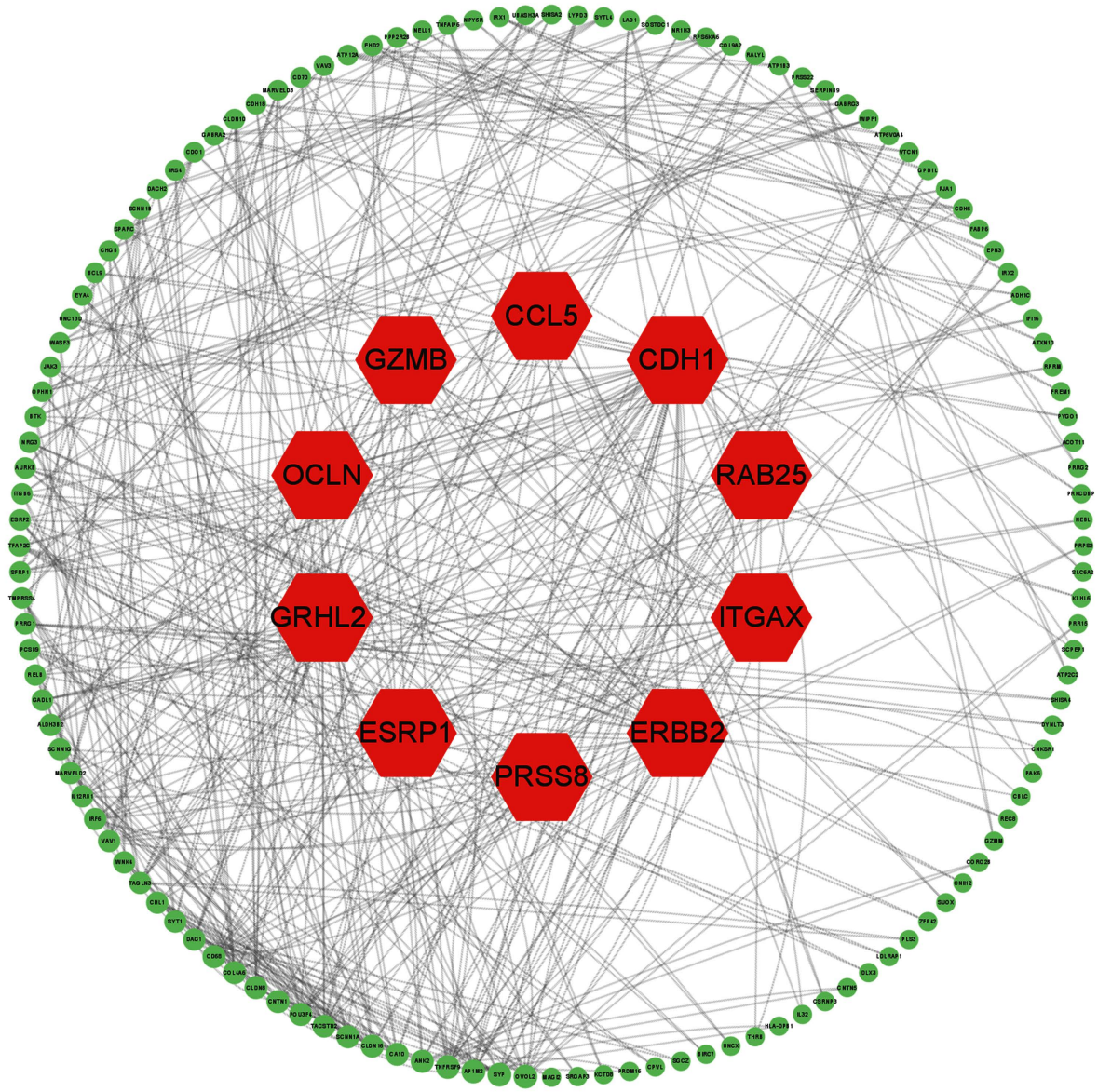
Table 5. Enrichment analysis of KIRP primary signature genes
表 5. KIRP 初步特征基因的功能富集

	Description	P value	Count
GO: 0034329	cell junction assembly	1.46E−05	8
GO: 1904861	excitatory synapse assembly	4.37E−05	3
GO: 0097106	postsynaptic density organization	6.10E−05	3
GO: 0099084	postsynaptic specialization organization	7.48E−05	3

对比肾细胞癌亚型特征基因和差异表达基因的富集分析，发现了 KIRC 在两种不同基因集的富集中都发现其与免疫系统的负调节，淋巴细胞的增殖和分化以及免疫细胞的增殖等相关通路有关(表 4 和表 S1)。这与 Xu 等[32] [33]在其研究中发现免疫反应是透明细胞肾细胞癌致癌作用和治疗效果的重要特征相一致。此外，一些研究[2] [34]也已证实 KIRC 比 KIRP 和 KICH 增加了免疫细胞浸润基因表达特征，并且有报告

[35]基于基因表达和回归分析等方法找到了 KIRC 的潜在免疫生物标志物，这些发现进一步支持了免疫反应在 KIRC 中的致癌作用。

免疫反应是肾透明细胞癌的致癌和治疗的重要特征。通过对肾透明细胞癌的特征基因进行蛋白质互作网络的分析(图 6)，得到了 10 个 hub 基因(*CDH1*、*ERBB2*、*ITGAX*、*ESRP1*、*GRHL2*、*GZMB*、*CCL5*、*OCLN*、*RAB25* 和 *PRSS8*)，其中 *ITGAX*、*CCL5* 和 *GZMB* 是表达量上调基因，其余是下调基因；对这几个基因进行了免疫细胞浸润性分析，发现 *ITGAX* 和 *CCL5* 表达量的升高与这六种免疫细胞浸润均显著相关($P < 0.05$) (图 7)，*GZMB* 的上调与 CD4 + T 细胞、CD8 + T 细胞和中性粒细胞的浸润性显著相关(图 S1)，*ERBB2*、*GRHL2* 和 *PRSS8* 的下调与 CD4 + T 细胞的浸润性显著相关(图 S1)。



(红色基因是 hub 基因，外环基因按照程度中心性从大(深色)到小(浅色)的顺序排列)

Figure 6. PPI network map of KIRC signature genes (Genes in red are hub genes, Outer ring genes are arranged in order of degree centrality from large (dark color) to small(light color))
图 6. KIRC 特征基因的 PPI 网络图

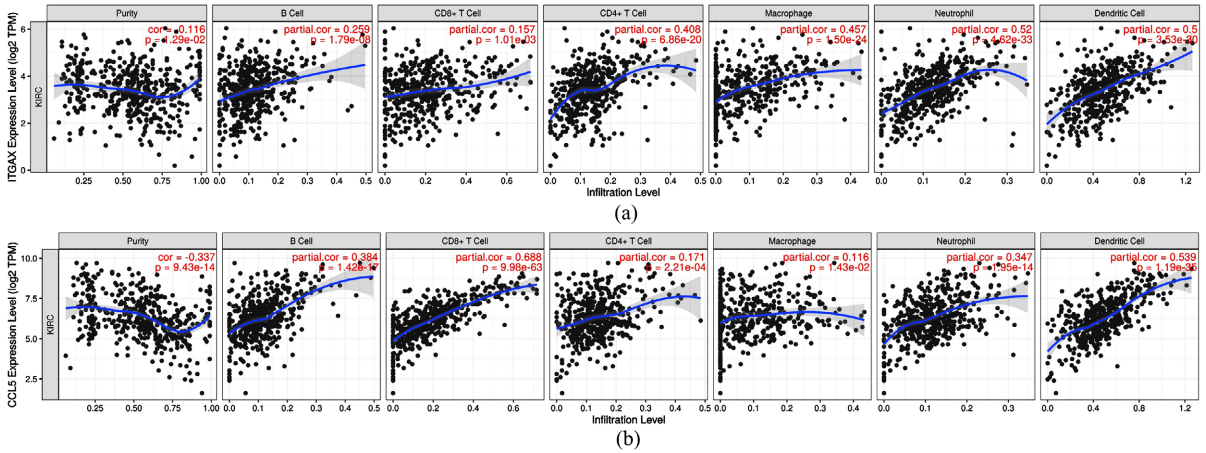


Figure 7. Immune infiltration of *ITGAX* and *CCL5*
图 7. *ITGAX* 和 *CCL5* 的免疫浸润

本研究通过构建 PPI 网络得到的 hub 基因中 *CCL5* 基因与 KIRC 的发生和发展显著相关。*CCL5* 是趋化因子 CC 家族的一员，CC 亚组的趋化因子在调节肿瘤微环境中起重要作用[36]。其主要表达于 T 细胞、巨噬细胞和部分肿瘤细胞[37]。*CCL5* 通过募集 T 细胞和巨噬细胞等间接促进肿瘤的增殖[36]。本研究发现 *CCL5* 基因的上调与 CD8 + T 细胞和树突状细胞的免疫浸润强相关，白等[38]的研究也证实了 *CCL5* 的上调在 KIRC 中与 CD4 + T 细胞、CD8 + T 细胞、Tregs 细胞和静息树突状细胞等有显著关系。有研究报告证实了 *CCL5* 基因可以作为 KIRC 患者的潜在治疗靶点[38] [39]。

4. 结论

本研究旨在从转录组数据和 DNA 甲基化数据两个组学层面对早期的 KIRC、KIRP 和 KICH 进行特征基因筛选及有监督的分类，寻找肾细胞癌亚型之间的不同分子特征。利用支持向量机、弹性网络和交叉验证等方法构建了三分类的分类器，筛选得到了贡献值最大的 6 个特征基因作为分类特征基因，准确性达到了 96.6%，精确性和敏感性分别是 93.4%、94.7%，F1-score 值和 Kappa 系数都达到了 0.94，说明了分类器的可靠性。同时，独立检验集正确率达到了 93.1%。需要说明的是，通过检索 GEO 数据库，得到的独立检验集仅包含了 KIRC 一种亚型，因此对 KIRP 和 KICH 需要在未来满足条件时加以验证。

对各个亚型中的特征基因使用 GO 分析进行了通路富集分析，发现 KIRC 主要与免疫细胞的增殖等相关通路有关，与其不同的是，KIRP 和 KICH 都和肾脏与肾脏系统的发育有关。此外，KIRP 还与钙离子的运输和泌尿生殖系统的发育有关，而 KICH 则与纤毛组织相关通路有关。本研究也发现 *CCL5* 的上调与免疫之间存在显著相关性，这也为精准免疫治疗提供了理论基础。

致　谢

感谢国家自然科学基金和重点研究开发项目的支持；此外，特别感谢 TCGA 数据库提供的原始数据。最后，特别感谢编辑和各位匿名审稿人对本文的支持和帮助。

基金项目

本研究由国家自然科学基金(61931013)和重点研究开发项目(2017YFC011104)资助。

参考文献

[1] Sung, H., Ferlay, J., Siegel, R.L., *et al.* (2021) Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and

Mortality Worldwide for 36 Cancers in 185 Countries. *CA*: *A Cancer Journal for Clinicians*, **71**, 209-249. https://doi.org/10.3322/caac.21660

[2] Chen, F., Zhang, Y., Senbabaoglu, Y., *et al.* (2016) Multilevel Genomics-Based Taxonomy of Renal Cell Carcinoma. *Cell Reports*, **14**, 2476-2489. https://doi.org/10.1016/j.celrep.2016.02.024

[3] Moch, H., Cubilla, A.L., Humphrey, P.A., *et al.* (2016) The 2016 WHO Classification of Tumours of the Urinary System and Male Genital Organs-Part A: Renal, Penile, and Testicular Tumours. *European Urology*, **70**, 93-105. https://doi.org/10.1016/j.eururo.2016.02.029

[4] Johnson, R. and Halder, G. (2014) The Two Faces of Hippo: Targeting the Hippo Pathway for Regenerative Medicine and Cancer Treatment. *Nature Reviews*. *Drug Discovery*, **13**, 63-79. https://doi.org/10.1038/nrd4161

[5] Shuch, B., Amin, A., Armstrong, A.J., *et al.* (2015) Understanding Pathologic Variants of Renal Cell Carcinoma: Distilling Therapeutic Opportunities from Biologic Complexity. *European Urology*, **67**, 85-97. https://doi.org/10.1016/j.eururo.2014.04.029

[6] Ricketts, C.J., De Cubas, A.A., Fan, H., *et al.* (2018) The Cancer Genome Atlas Comprehensive Molecular Characterization of Renal Cell Carcinoma. *Cell Reports*, **23**, 3698. https://doi.org/10.1016/j.celrep.2018.06.032

[7] Garje, R., Elhag, D., Yasin, H.A., *et al.* (2021) Comprehensive Review of Chromophobe Renal Cell Carcinoma. *Critical Reviews in Oncology/Hematology*, **160**, Article ID: 103287. https://doi.org/10.1016/j.critrevonc.2021.103287

[8] Sun, M., Tong, P., Kong, W., *et al.* (2017) HNF1B Loss Exacerbates the Development of Chromophobe Renal Cell Carcinomas. *Cancer Research*, **77**, 5313-5326. https://doi.org/10.1158/0008-5472.CAN-17-0986

[9] Davis, C.F., Ricketts, C.J., Wang, M., *et al.* (2014) The Somatic Genomic Landscape of Chromophobe Renal Cell Carcinoma. *Cancer Cell*, **26**, 319-330. https://doi.org/10.1016/j.ccr.2014.07.014

[10] Cancer Genome Atlas Research N, Linehan, W.M., Spellman, P.T., *et al.* (2016) Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. *The New England Journal of Medicine*, **374**, 135-145. https://doi.org/10.1056/NEJMoa1505917

[11] Han, S., Hwang, S.I. and Lee, H.J. (2019) The Classification of Renal Cancer in 3-Phase CT Images Using a Deep Learning Method. *Journal of Digital Imaging*, **32**, 638-643. https://doi.org/10.1007/s10278-019-00230-2

[12] Chaudry, Q., Raza, S.H., Sharma, Y., *et al.* (2008) Improving Renal Cell Carcinoma Classification by Automatic Region of Interest Selection. *Proceedings IEEE International Symposium on Bioinformatics and Bioengineering*, Athens, 8-10 October 2008, 1-6. https://doi.org/10.1109/BIBE.2008.4696796

[13] Gui, C.P., Wei, J.H., Chen, Y.H., *et al.* (2021) A New Thinking: Extended Application of Genomic Selection to Screen Multiomics Data for Development of Novel Hypoxia-Immune Biomarkers and Target Therapy of Clear Cell Renal Cell Carcinoma. *Briefings in Bioinformatics*, **22**, bbab173. https://doi.org/10.1093/bib/bbab173

[14] Deng, R., Li, J., Zhao, H., *et al.* (2021) Identification of Potential Biomarkers Associated with Immune Infiltration in Papillary Renal Cell Carcinoma. *Journal of Clinical Laboratory Analysis*, **35**, e24022. https://doi.org/10.1002/jcla.24022

[15] Tomczak, K., Czerwinska, P. and Wiznerowicz, M. (2015) The Cancer Genome Atlas (TCGA): An Immeasurable Source of Knowledge. *Contemporary Oncology* (*Poznan*, *Poland*), **19**, A68-A77. https://doi.org/10.5114/wo.2014.47136

[16] Kananen, L., Marttila, S., Nevalainen, T., *et al.* (2016) Aging-Associated DNA Methylation Changes in Middle-Aged Individuals: The Young Finns Study. *BMC Genomics*, **17**, 103. https://doi.org/10.1186/s12864-016-2421-z

[17] Yang, X., Gao, L. and Zhang, S. (2017) Comparative Pan-Cancer DNA Methylation Analysis Reveals Cancer Common and Specific Patterns. *Briefings in Bioinformatics*, **18**, 761-773. https://doi.org/10.1093/bib/bbw063

[18] Palou-Marquez, G., Subirana, I., Nonell, L., *et al.* (2021) DNA Methylation and Gene Expression Integration in Cardiovascular Disease. *Clinical Epigenetics*, **13**, 75. https://doi.org/10.1186/s13148-021-01064-y

[19] Eden, S.K., Li, C. and Shepherd, B.E. (2022) Nonparametric Estimation of Spearman's Rank Correlation with Bivariate Survival Data. *Biometrics*, **78**, 421-434. https://doi.org/10.1111/biom.13453

[20] Hazra, A. and Gogtay, N. (2016) Biostatistics Series Module 6: Correlation and Linear Regression. *Indian Journal of Dermatology*, **61**, 593-601. https://doi.org/10.4103/0019-5154.193662

[21] Mishra, P., Singh, U., Pandey, C.M., *et al.* (2019) Application of Student's t-Test, Analysis of Variance, and Covariance. *Annals of Cardiac Anaesthesia*, **22**, 407-411. https://doi.org/10.4103/aca.ACA_94_19

[22] 温建鑫, 王学栋, 李晓琴, 等. 乳腺癌发生的特征基因筛选及模式识别[J]. 生物化学与生物物理进展, 2017, 44(11): 1016-1025.

[23] Tong, Y., Yu, Y., Zheng, H., *et al.* (2021) Differentially Expressed Genes in Clear Cell Renal Cell Carcinoma as a Potential Marker for Prognostic and Immune Signatures. *Frontiers in Oncology*, **11**, Article ID: 776824. https://doi.org/10.3389/fonc.2021.776824

[24] Lasseigne, B.N. and Brooks, J.D. (2018) The Role of DNA Methylation in Renal Cell Carcinoma. *Molecular Diagnosis & Therapy*, **22**, 431-442. https://doi.org/10.1007/s40291-018-0337-9

[25] Wang, X., Shang, W., Chang, Y., *et al.* (2018) Methylation Signature Genes Identification of the Lung Squamous Cell Carcinoma Occurrence and Recognition Research. *Journal of Computational Biology*: *A Journal of Computational Molecular Cell Biology*, **25**, 1161-1169. https://doi.org/10.1089/cmb.2018.0069

[26] Lopez-Blanco, J.R. and Chacon, P. (2016) New Generation of Elastic Network Models. *Current Opinion in Structural Biology*, **37**, 46-53. https://doi.org/10.1016/j.sbi.2015.11.013

[27] Li, Q., Wen, Z. and He, B. (2020) Adaptive Kernel Value Caching for SVM Training. *IEEE Transactions on Neural Networks and Learning Systems*, **31**, 2376-2386.

[28] Faris, H., Habib, M., Faris, M., *et al.* (2020) Medical Speciality Classification System Based on Binary Particle Swarms and Ensemble of One vs. Rest Support Vector Machines. *Journal of Biomedical Informatics*, **109**, Article ID: 103525. https://doi.org/10.1016/j.jbi.2020.103525

[29] Roskoski Jr., R. (2020) The Role of Fibroblast Growth Factor Receptor (FGFR) Protein-Tyrosine Kinase Inhibitors in the Treatment of Cancers Including Those of the Urinary Bladder. *Pharmacological Research*, **151**, Article ID: 104567. https://doi.org/10.1016/j.phrs.2019.104567

[30] Cunningham, J.T., Moreno, M.V., Lodi, A., *et al.* (2014) Protein and Nucleotide Biosynthesis Are Coupled by a Single Rate-Limiting Enzyme, PRPS2, to Drive Cancer. *Cell*, **157**, 1088-1103. https://doi.org/10.1016/j.cell.2014.03.052

[31] Finan, C., Gaulton, A., Kruger, F.A., *et al.* (2017) The Druggable Genome and Support for Target Identification and Validation in Drug Development. *Science Translational Medicine*, **9**, eaag1166. https://doi.org/10.1126/scitranslmed.aag1166

[32] Xu, W.H., Xu, Y., Wang, J., *et al.* (2019) Prognostic Value and Immune Infiltration of Novel Signatures in Clear Cell Renal Cell Carcinoma Microenvironment. *Aging*, **11**, 6999-7020. https://doi.org/10.18632/aging.102233

[33] Zhan, X., Liu, Y., Yu, C.Y., *et al.* (2020) A Pan-Kidney Cancer Study Identifies Subtype Specific Perturbations on Pathways with Potential Drivers in Renal Cell Carcinoma. *BMC Medical Genomics*, **13**, 190. https://doi.org/10.1186/s12920-020-00827-5

[34] Geissler, K., Fornara, P., Lautenschlager, C., *et al.* (2015) Immune Signature of Tumor Infiltrating Immune Cells in Renal Cancer. *Oncoimmunology*, **4**, e985082. https://doi.org/10.4161/2162402X.2014.985082

[35] Xiang, Z., Shen, E., Li, M., *et al.* (2021) Potential Prognostic Biomarkers Related to Immunity in Clear Cell Renal Cell Carcinoma Using Bioinformatic Strategy. *Bioengineered*, **12**, 1773-1790. https://doi.org/10.1080/21655979.2021.1924546

[36] Aldinucci, D. and Casagrande, N. (2018) Inhibition of the CCL5/CCR5 Axis against the Progression of Gastric Cancer. *International Journal of Molecular Sciences*, **19**, 1477. https://doi.org/10.3390/ijms19051477

[37] Leighton, S.P., Nerurkar, L., Krishnadas, R., *et al.* (2018) Chemokines in Depression in Health and in Inflammatory Illness: A Systematic Review and Meta-Analysis. *Molecular Psychiatry*, **23**, 48-58. https://doi.org/10.1038/mp.2017.205

[38] Bai, S., Wu, Y., Yan, Y., *et al.* (2020) The Effect of CCL5 on the Immune Cells Infiltration and the Prognosis of Patients with Kidney Renal Clear Cell Carcinoma. *International Journal of Medical Sciences*, **17**, 2917-2925. https://doi.org/10.7150/ijms.51126

[39] Walens, A., Dimarco, A.V., Lupo, R., *et al.* (2019) CCL5 Promotes Breast Cancer Recurrence through Macrophage Recruitment in Residual Tumors. *Elife*, **8**, e43653. https://doi.org/10.7554/eLife.43653

附 录

Table S1. Functional enrichment of KIRC differentially expressed genes
表 S1. KIRC 差异表达基因的功能富集

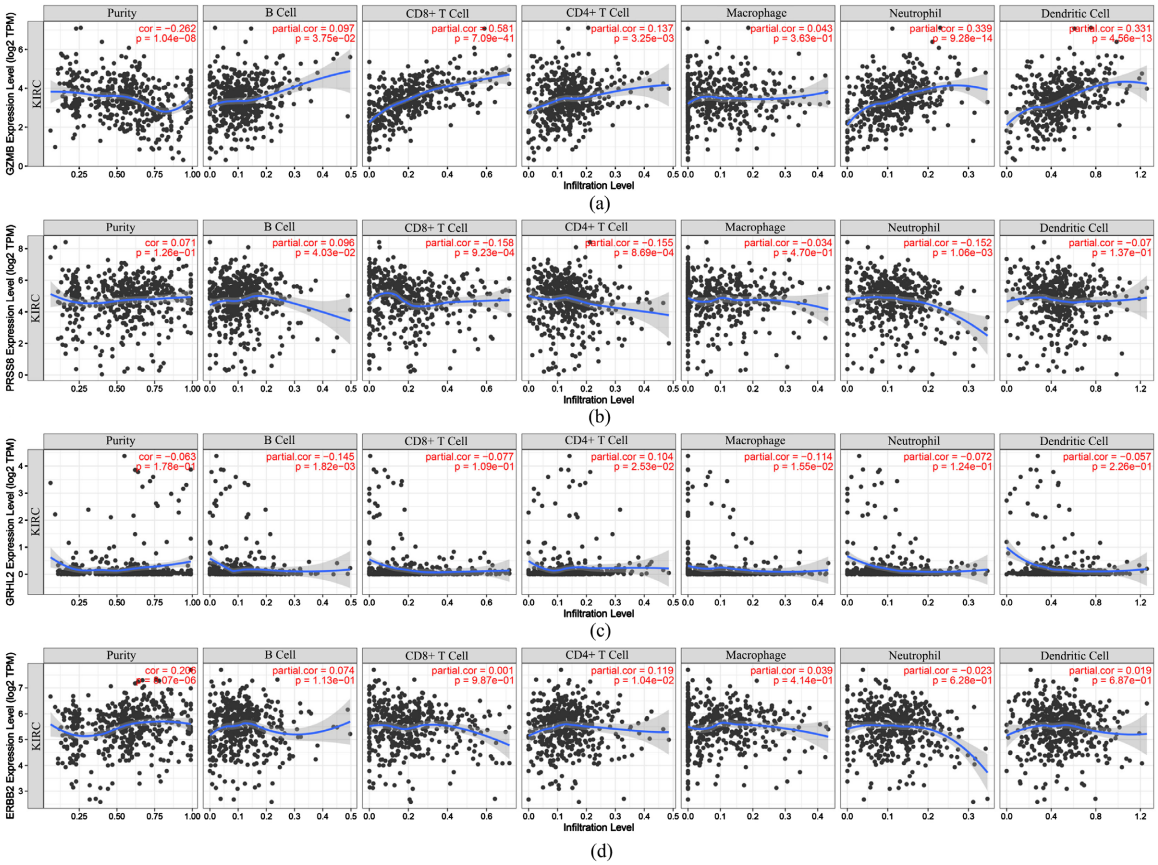
	Description	P value	Count
GO: 0042110	T cell activation	1.50E−18	114
GO: 0002683	negative regulation of immune system process	4.47E−12	88
GO: 1903131	mononuclear cell differentiation	1.37E−11	88
GO: 0007159	leukocyte cell-cell adhesion	8.40E−13	84
GO: 0050863	regulation of T cell activation	2.38E−13	79
GO: 0030098	lymphocyte differentiation	8.29E−11	78
GO: 0070661	leukocyte proliferation	4.33E−13	76
GO: 1903037	regulation of leukocyte cell-cell adhesion	2.34E−11	75
GO: 0046651	lymphocyte proliferation	1.50E−13	72
GO: 0032943	mononuclear cell proliferation	2.64E−13	72
GO: 0070663	regulation of leukocyte proliferation	5.16E−12	62
GO: 0050670	regulation of lymphocyte proliferation	9.53E−13	60
GO: 0032944	regulation of mononuclear cell proliferation	1.45E−12	60
GO: 0002703	regulation of leukocyte mediated immunity	8.35E−13	58
GO: 0042098	T cell proliferation	1.59E−12	55

Table S2. Functional enrichment of KIRP differentially expressed genes
表 S2. KIRP 差异表达基因的功能富集

	Description	P value	Count
GO: 0030001	metal ion transport	4.61E−13	100
GO: 0003012	muscle system process	5.04E−11	90
GO: 0090066	regulation of anatomical structure size	9.60E−09	89
GO: 0006816	calcium ion transport	1.56E−11	85
GO: 0001655	urogenital system development	1.79E−13	76
GO: 0006936	muscle contraction	1.87E−10	74
GO: 0072001	renal system development	1.60E−13	71
GO: 0001822	kidney development	3.61E−13	69
GO: 1903522	regulation of blood circulation	1.06E−09	63
GO: 0003018	vascular process in circulatory system	1.64E−12	62
GO: 0070588	calcium ion transmembrane transport	1.85E−08	62
GO: 0060047	heart contraction	2.60E−08	58
GO: 0098657	import into cell	2.56E−09	53
GO: 0008016	regulation of heart contraction	2.56E−08	53
GO: 0072006	nephron development	6.22E−13	44

Table S3. Functional enrichment of KICH differentially expressed genes
表 S3. KICH 差异表达基因的功能富集

	Description	P value	Count
GO: 0044782	cilium organization	2.54E−21	104
GO: 0060271	cilium assembly	1.82E−20	99
GO: 0007018	microtubule-based movement	5.34E−13	81
GO: 0007389	pattern specification process	2.06E−09	80
GO: 0001655	urogenital system development	4.17E−10	67
GO: 0072001	renal system development	5.85E−10	62
GO: 0001822	kidney development	3.57E−09	59
GO: 0003341	cilium movement	2.41E−12	47
GO: 0072006	nephron development	3.88E−09	37
GO: 0001539	cilium or flagellum-dependent cell motility	4.64E−09	35
GO: 0060285	cilium-dependent cell motility	4.64E−09	35
GO: 0001578	microtubule bundle formation	1.20E−10	34
GO: 0072073	kidney epithelium development	3.96E−08	34
GO: 0035082	axoneme assembly	4.33E−13	31
GO: 0072009	nephron epithelium development	7.44E−09	31



(a)~(d)依次是 GZMB、PRSS8、GRHL2 和 ERBB2

Figure S1. Immune infiltration of hub gene
图 S1. hub 基因的免疫浸润