doi:10.7659/j.issn.1005-6947.2022.10.011
http://dx.doi.org/10.7659/j.issn.1005-6947.2022.10.011
Chinese Journal of General Surgery, 2022, 31(10):1355-1362.

·基础研究·

基于机器学习的结直肠癌微卫星不稳定基因挖掘及其应用价值分析

李秀勤[1]，韩腾辉[2]，王帅[3]，沈刚[3]，朱军[1]

（1. 中国人民解放军南部战区空军医院 普通外科，广东 广州 510000；2. 中国人民解放军空军军医大学西京医院 神经内科，陕西 西安 710000；3. 中国人民解放军空军西安飞行学院一旅明港场站医院 门诊部，河南 信阳 463200）

摘 要

背景与目的：结直肠癌（CRC）是全球第三大最常诊断的恶性肿瘤和第二大癌症死亡原因。最新指南推荐所有的CRC患者需要进行微卫星不稳定（MSI）的检测。MSI患者往往具有错配修复蛋白缺失（dMMR）。MSI/dMMR状态已被用作生物标志物预测对免疫治疗的有利反应和预后。然而MSI特征基因及其与肿瘤浸润的免疫细胞的关系未进行阐述。因此本研究通过使用机器学习的方式发掘CRC中新型的MSI特征基因，并且验证其的诊断价值及其与免疫细胞浸润的关系。

方法：按照纳入排除标准，将GEO数据库中GSE39582数据集作为训练集，将TCGA数据库中COAD数据集作为外部验证集。使用机器学习的方法（LASSO回归、SVM-RFE算法），在GSE39582结直肠癌数据集中筛选MSI特征基因，并在TCGA结直肠癌数据中进行验证。采用受试者工作特征（ROC）曲线和曲线下面积（AUC）评价基因对MSI的诊断效能。CIBERSORT算法评估肿瘤样本浸润的免疫细胞成分，Spearman相关性分析验证MSI特征基因和免疫细胞的关系。

结果：训练集共纳入536例CRC患者，其中高频MSI（MSI-H）77例（14.37%）。在验证集中，共计389例CRC患者，其中MSI-H 67例（17.22%）。基线资料分析显示，MSI-H/dMMR CRC的TNM分期存活率优于低频MSI（MSI-L）或微卫星稳定（MSS）/错配蛋白完整（pMMR）CRC（$P<0.05$）。在GSE39582数据集中，LASSO回归筛选MSI特征基因21个，SVM-RFE算法筛选基因6个，结合两种算法确定MSI特征基因为*EIF5A*、*CXCL13*、*HNRNPL*、*HOXC6*、*RPL22L1*、*Y16709*。在TCGA数据库中进一步验证MSI特征基因的诊断效能，研究发现*EIF5A*的诊断效能最高。在训练集和验证集中，*EIF5A*的AUC值分别为0.922和0.805。同时，Spearman相关性分析发现，*EIF5A*主要与$CD8^+$T细胞，活化的树突状细胞，辅助性T细胞，M1型巨噬细胞，γδT细胞，中性粒细胞成正相关；与$CD4^+$记忆性T细胞，M2型巨噬细胞，静止树突状细胞，嗜酸性粒细胞，调节性T细胞呈负相关。

结论：CRC的新型MSI特征基因分析结果表明，*EIF5A*对CRC MSI的诊断具有较好的诊断作用和临床价值，同时提示*EIF5A*与免疫细胞及免疫微环境相关。因此，*EIF5A*可能成为免疫检查点治疗的新型标志物。

关键词 结直肠肿瘤；微卫星不稳定性；DNA错配修复；淋巴细胞，肿瘤浸润；机器学习

中图分类号：R735.3

基金项目：国家自然科学基金资助项目（82100680）。

收稿日期：2021-11-29；**修订日期：**2022-04-18。

作者简介：李秀勤，中国人民解放军南部战区空军医院主治医师，主要从事消化道肿瘤临床方面的研究。

通信作者：朱军，Email: zjsty@fmmu.edu.cn

Mining of genes involved in microsatellite instability in colorectal cancer through machine learning and evaluation of their application values

LI Xiuqin[1], HAN Tenghui[2], WANG Shuai[3], SHEN Gang[3], ZHU Jun[1]

(1. Department of General Surgery, the Southern Theater Air Force Hospital, Guangzhou 510000, China; 2. Department of Neurology, Air Force Medical University, Xi'an 710000, China; 3. Ming Gang Station Hospital, Xi'an Institute of Flight of the Air Force, Xingyang, Henan 463200, China)

Abstract

Background and Aims: Colorectal cancer (CRC) is the third most commonly diagnosed malignancy and the second leading cause of cancer death worldwide. The latest guidelines recommend that all CRC patients need to be tested for microsatellite instability (MSI). MSI patients often have deficient mismatch repair (dMMR). The MSI/dMMR has been used as a biomarker for predicting the favorable response to immunotherapy and prognosis of patients. However, MSI signature genes and their relationship to tumor-infiltrating immune cells have not been fully described. Therefore, this study was conducted to discover novel MSI signature genes in CRC through machine learning and verify their diagnostic values and relationships with immune cell infiltration.

Methods: According to the inclusion and exclusion criteria, the GSE39582 dataset in GEO database was used as the training set, and the COAD dataset in TCGA database was used as the external validation set. Using machine learning methods (LASSO regression and SVM-RFE algorithm), MSI signature genes were screened in the GSE39582 CRC data set and validated in the TCGA COAD dataset. Receiver operating characteristic (ROC) curve and area under the curve (AUC) were used to evaluate the diagnostic performance of genes for MSI. The CIBERSORT algorithm evaluated each sample's immune infiltrating cell components, and Spearman correlation analysis was used to verify the relationship between MSI signature genes and immune cells.

Results: A total of 536 CRC patients were included in training set, of which 77 cases (for 14.37%) were high microsatellite instability (MSI-H). In validation set, there were a total of 389 CRC patients, of which 67 cases (17.22%) were MSI-H. The baseline data analysis showed that the TNM profiles and survival rates in MSI-H/dMMR CRC were superior to those in low microsatellite instability (MSI-L) or microsatellite stable (MSS)/proficient mismatch repair (pMMR) CRC (P<0.05). In GSE39582 dataset, 21 MSI signature genes were screened by LASSO regression, and 6 genes were screened by SVM-RFE algorithm. The MSI signature genes were identified as *EIF5A*, *CXCL13*, *HNRNPL*, *HOXC6*, *RPL22L1*, and *Y16709* by combining the two algorithms. The diagnostic efficacy of MSI signature genes was further verified in TCGA database, and *EIF5A* was found to have the highest diagnostic efficacy. The AUC values for *EIF5A* in training and validation sets were 0.922 and 0.805, respectively. At the same time, Spearman correlation analysis found that *EIF5A* was mainly positively correlated with $CD8^+$ T cells, activated dendritic cells, helper T cells, M1 macrophages, γδ T cells, and neutrophils; it was negatively correlated with $CD4^+$ memory T cells, M2 macrophages, quiescent dendritic cells, eosinophils, and regulatory T cells.

Conclusion: Analysis of novel MSI signature genes in CRC shows that *EIF5A* has a good diagnostic performance and clinical value for CRC MSI status. It is also associated with immune cells and immune microenvironment. Thus, *EIF5A* may become a new marker for immune checkpoint therapy.

Key words　Colorectal Neoplasms; Microsatellite Instability; DNA Mismatch Repair; Lymphocytes, Tumor-Infiltrating; Machine Learning

CLC number: R735.3

流行病学研究表明，结直肠癌（colorectal cancer，CRC）作为全球发病率排第 2 位、致死率排第 3 位的肿瘤，已经成为威胁人类健康的重要疾病之一[1]。目前，我国结直肠癌的发病率在常见恶性肿瘤中排第 3 位、病死率排第 5 位[2]。各大指南推荐所有的 CRC 患者需要检测微卫星状态，以完善患者的临床诊断，治疗指导和预后评估[3]。微卫星广泛存在于原核及真核生物基因组中，具有较高的遗传稳定性，但在错配修复基因功能发生异常时，子代细胞微卫星的重复核苷酸数量可以增多或减少，从而导致微卫星的长度不再保持一致，这种现象称微卫星不稳定（microsatellite instability，MSI）。而一种或多种错配修复蛋白的缺失（deficient mismatch repair，dMMR）往往会导致高频 MSI（MSI-H）[4]。此外，有文献[5]报道错配修复蛋白的免疫组化检测与 MSI 的 PCR 检测结果具有高度的一致性，因此，本研究将 MSI-H/dMMR 作为相似的一组进行分析。随着免疫治疗的兴起，尤其是抗 PD-1/PD-L1 药物的实体肿瘤的治疗出现了新的转机。PD-1/PD-L1 是肿瘤细胞进行免疫逃逸的重要分子通路，抗 PD-1/PD-L1 可以明显增强肿瘤浸润免疫细胞的杀伤能力[6]。但是，由于 PD-1/PD-L1 的表达量低而使得 CRC 的免疫治疗受到了极大的限制。目前，CRC 的免疫治疗现在主要适用于 MSI-H/dMMR 的患者[7-8]。目前 MSI-H 和免疫检查点治疗的内在机制研究开展较少，而除了错配修复蛋白基因的研究外，关于结直肠癌 MSI 特征基因也未得到深入广泛的研究。本研究的目的是使用机器学习筛选 MSI 诊断效率最高的基因，并且研究基因和肿瘤免疫微环境的相关性。

随着机器学习和人工智能在医学领域的广泛运用，病理辅助诊断[9]、疾病精确诊断[10]和个性化治疗[11]已逐渐在临床上得以实现。同时二代测序技术为 CRC 患者精准治疗提供了极大的便利。因此，本研究的主要目的是，使用机器学习和数据库验证分析等方法在 CRC 患者测序数据中，发掘新型的 MSI 特征基因，为临床研究和应用提供新的线索和方向。

1　资料与方法

1.1　研究对象

研究对象为临床确诊的 CRC 患者人群。研究对象的纳入标准为：(1) 年龄≥18 周岁；(2) 已知微卫星状态或错配修复基因缺失情况的患者；(3) 测序数据完整的患者（基因二代测序或者组织芯片）。排除标准为：(1) 合并其他肿瘤的患者；(2) 生存时间少于 30 d 的患者；(3) 未采取手术治疗而无法获取大体病理资料的患者。该研究已通过中国人民解放军南部战区空军医院审核批准。

1.2　数据收集

在 GEO 官网（https://www.ncbi.nlm.nih.gov/geo）下载 CRC 完整测序数据 GSE39582，在 TCGA 官网（https://portal.gdc.cancer.gov）下载 CRC 测序数据 TCGA-COAD。由于 GSE39582 的 CRC 样本量较大，因此本研究将 GSE39582 作为训练集，将 TCGA-COAD 作为外部验证集。此外，使用 Linear Models for Microarray Data（LIMMA）包中 normalizeBetweenArrays 函数对数据进行标准化处理，通过 SVA 包中的 Combat 函数去除 2 个数据集的批次效应。

1.3　研究对象分组

按照微卫星状态或者错配修复基因的表达水平，本研究分别将 GSE39582 和 TCGA-COAD 数据集中的患者分为 MSI-H/dMMR 组和低频度 MSI（MSI-L）或微卫星稳定（MSS）/错配蛋白完整（pMMR）组。本研究把 MSI-L-MSS/pMMR 组作为对照组，MSI-H/dMMR 组作为观察组。使用 LIMMA 包对差异基因进行筛选，其校正方法为 FDR 法。筛选条件为：|log2（差异倍数）|>1.5 并且 FDR 值<0.05。

1.4　机器算法

为了得到更精确的 MSI 特征基因，分别使用 LASSO 回归算法和支持向量机－递归特征消除（SVM-RFE）算法对上述得到的差异基因进行筛选。LASSO 回归算法：使用 glmnet 包，alpha 参数设置为 1，交叉验证为 10，高斯分布用于交叉验证的损失。LASSO 筛选的基因定义为：二项式误差最小

　

值时对应的基因数目。SVM-RFE算法：使用caret，kernlab，e1071包，对模型进行内部交叉验证，采用的方法为“svmRadial”，最后筛选的基因为交叉验证误差（RSME）最小值的基因数目。受试者工作特征曲线（ROC）用以评价MSI特征基因的诊断效能，曲线下面积（AUC）值为MSI特征基因的评价指标。以抽样的方式计算AUC值95%可信区间，抽样方法为bootstrap法。

1.5 肿瘤浸润的免疫细胞评估

CIBERSORT算法[12]评估GSE39582测序数据的免疫细胞浸润情况，P<0.05作为预测准确的筛选标准。评估的免疫细胞主要包含：CD4+T细胞，CD8+T细胞，树突状细胞，辅助型T细胞，M1巨噬细胞，M2型巨噬细胞，M0型巨噬细胞，中性粒细胞，B细胞，记忆性B细胞和肥大细胞。本研究使用相关性分析研究MSI特征基因与肿瘤浸润免疫细胞的关联性，以探索MSI特征基因对肿瘤免疫微环境的影响。同时，MSI特征基因与免疫细胞的相关性也进行了分析研究。

1.6 统计学处理

计量资料中，符合正态分布方差齐性的数据以平均数±标准差（$\bar{x} \pm s$）的方式来表示，其检验方式为Student't检验或方差分析；不符合正态分布或者方差齐性的数据使用中位数（四分位间距）[M（IQR）]，检验方式为非参数检验。计数资料，表达方式为例数（百分数）[n（%）]，其检验方式为χ^2检验或Fisher精确概率。特征基因与免疫细胞的相关性分析采用的是Spearman秩相关分析。本研究中使用的其余R包有：dplyr，ggplot2，pROC等等。P<0.05为差异有统计学意义。

2 结 果

2.1 基线资料特征

GSE39582数据中共收集536例CRC患者，其中MSI-H患者77例；MSI-L/MSS患者459例。MSI-H组55例存活，存活率为71.4%，MSI-L/MSS组299例存活，存活率为65.1%，MSI-H组的存活率高于MSI-L/MSS组（P=0.001）。在TNM分期系统中，MSI-H组的患者均早于MSI-L/MSS组患者（T分期：P=0.036；N分期：P=0.007；M分期：P=0.02）。患者年龄，性别和生存时间在MSI-H组和MSI-L/MSS组的差异无统计学意义（均P>0.05）。

TCGA-COAD数据中共收集389例CRC患者，其中MSI-H患者67例；MSI-L/MSS患者322例。MSI-H组57例存活，存活率为85.1%，MSI-L/MSS组258例存活，存活率为80.1%，两组的存活率差异无统计学意义（P=0.442）。在N分期和M分期中，MSI-H组的患者早于MSI-L/MSS组患者（N分期：P<0.001；M分期：P=0.014）。T分期，患者年龄，性别和生存时间在两组中的差异无统计学意义（均P>0.05）（表1）。

表1 TCGA和GEO数据集的基线资料特征

Table 1 Baseline features of CRC patients in TCGA and GEO datasets

项目	TCGA数据集		P	GEO数据集		P
	MSI-H(n=67)	MSI-L/MSS(n=322)		MSI-H(n=77)	MSI-L/MSS(n=459)	
年龄(岁,$\bar{x} \pm s$)	68.8±13.8	65.9±12.4	0.115	68.9±16.3	66.9±12.7	0.296
性别[n(%)]						
男	38(56.7)	144(44.7)	0.098	38(49.4)	261(56.9)	0.269
女	29(43.3)	178(55.3)		39(50.6)	198(43.1)	
生存状态[n(%)]						
存活	57(85.1)	258(80.1)	0.442	55(71.4)	299(65.1)	0.001
死亡	10(14.9)	64(19.9)		19(24.7)	160(34.9)	
未知	—	—		3(3.9)	0(0.0)	
生存时间(月)	51.8(26.7~86.2)	41.3(26.2~86.0)	0.402	47.0(26.2~70.2)	52.0(26.0~81.0)	0.148
T分期[n(%)]						
T0	—	—	0.289	1(1.3)	3(0.7)	0.036
T1	3(4.5)	5(1.6)		1(1.3)	11(2.4)	
T2	13(19.4)	56(17.4)		10(12.9)	39(8.5)	
T3	46(68.6)	222(68.9)		43(55.9)	305(66.4)	
T4	5(7.5)	39(12.1)		22(28.6)	81(17.6)	
TX	—	—		0(0.00)	20(4.4)	

表 1　TCGA 和 GEO 数据集的基线资料特征（续）

Table 1　Baseline features of CRC patients in TCGA and GEO datasets (continued)

项目	TCGA 数据集		*P*	GEO 数据集		*P*
	MSI-H（*n*=67）	MSI-L/MSS（*n*=322）		MSI-H（*n*=77）	MSI-L/MSS（*n*=459）	
N 分期[*n*（%）]						
N0	54（80.6）	174（54.0）	<0.001	50（64.9）	220（47.9）	0.007
N1	10（14.9）	84（26.1）		15（19.5）	120（26.1）	
N2	3（4.5）	64（19.9）		12（15.6）	87（19.0）	
NX	—	—		0（0.00）	32（7.0）	
M 分期[*n*（%）]						
M0	59（88.1）	242（75.2）	0.014	72（93.5）	383（83.4）	0.02
M1	2（2.99）	54（16.8）		2（2.60）	56（12.2）	
MX	6（8.96）	26（8.07）		3（3.90）	20（4.36）	

2.2　差异基因分析

为了全面筛选 MSI 特征基因，本研究首先按照预先设置的分组情况，使用 LIMMA 包对每个测序基因进行筛选。在 GSE39582 数据中，差异基因分析结果如图 1 所示：MSI-H 组 17 个基因上调（红色点），17 个基因下调（绿色点），差异具有统计学意义。

2.3　LASSO 回归和 SVM 筛选 MSI 特征基因

为了进一步筛选相关基因，使用两种机器学习的方式对差异基因进行探究。在 LASSO 回归中，21 个差异基因在模型中被保留下来（图 2A）。在 SVM-RFE 分析中，6 个差异基因被确定（图 2B）。两种机器算法确定的基因在取交集后，最后得到 6 个 MSI 特征基因：*EIF5A*、*CXCL13*、*HNRNPL*、*HOXC*、*RPL22L1*、*Y16709*。

ROC 曲线验证 MSI 特征基因的诊断效能，结果详见表 2。在训练集（GSE39582）中，6 个基因的 AUC 值都在 0.75 以上，其中 *EIF5A*，*HNRNPL* 和 *Y16709* 的 AUC 值达 0.95 以上。在验证集（TCGA-COAD）中，*EIF5A* 的诊断效能最高（AUC=0.805）而 *Y16709* 基因在 TCGA 数据未发现。因此，本研究最终将 *EIF5A* 作为 MSI 的特征基因。

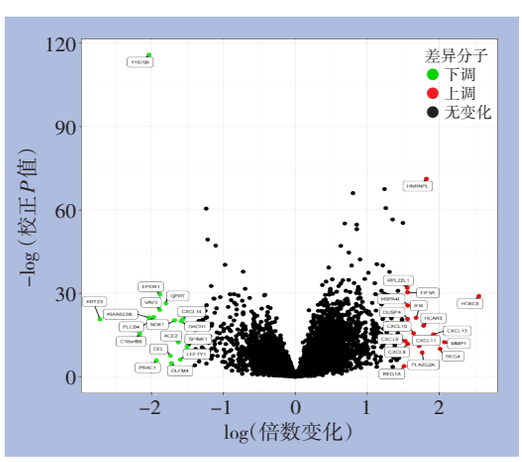

图 1　MSI 差异性基因的火山图

Figure 1　Volcano diagram of differentially expressed genes of MSI

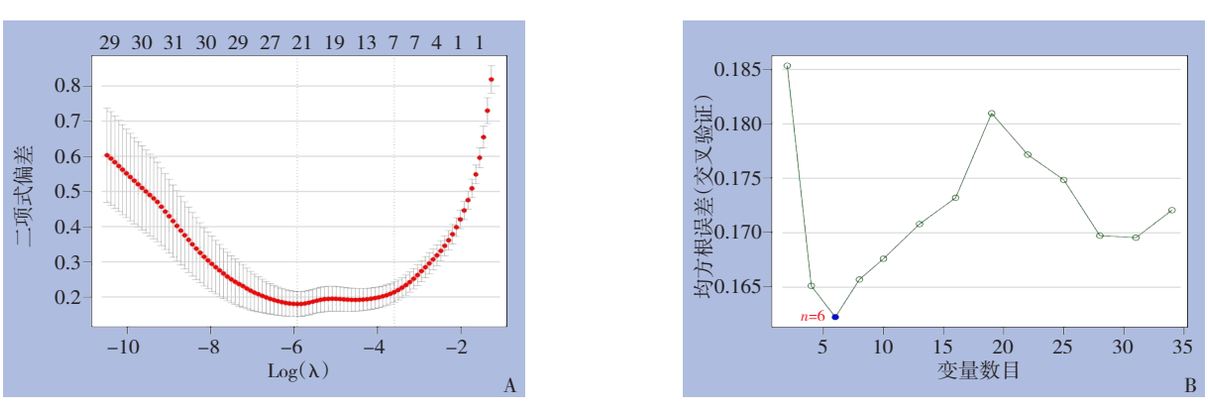

图 2　LASSO 回归和 SVM-RFE 筛选特征基因　A：LASSO 回归筛选特征基因的过程；B：SVM-RFE 中误差与变量数目的关系

Figure 2　MSI-related genes identified by LASSO regression and SVM-RFE methods　A: Selection of MSI-related genes by LASSO regression; B: The relationship between error and number of genes in SVM-RFE

表2 不同基因对CRC MSI状态的诊断效能

Table 2 Diagnostic efficacy of different genes for MSI status in colorectal cancer

基因	GSE39582(训练集)	TCGA-COAD(验证集)
EIF5A	0.922(0.887~0.952)	0.805(0.726~0.872)
CXCL13	0.765(0.709~0.819)	0.756(0.694~0.815)
HNRNPL	0.952(0.922~0.979)	0.518(0.441~0.598)
HOXC6	0.829(0.776~0.876)	0.780(0.709~0.846)
RPL22L1	0.895(0.851~0.932)	0.786(0.704~0.858)
Y16709	0.986(0.978~0.993)	—

2.4 *EIF5A*基因与肿瘤浸润免疫细胞的关系

CIBERSORT算法解析GSE39582的肿瘤免疫细胞浸润情况。在计算每种免疫细胞的评分之后，我们分析MSI特征基因*EIF5A*与免疫细胞的相关性。图3显示：$CD8^+$T细胞，活化的树突状细胞，辅助性T细胞，M1型巨噬细胞，γδT细胞，中性粒细胞与EIF5A成正相关（均$P<0.05$）；$CD4^+$记忆性T细胞，M2型巨噬细胞，静止树突状细胞，嗜酸性粒细胞，调节性T细胞（Treg）与*EIF5A*呈负相关（均$P<0.05$）。

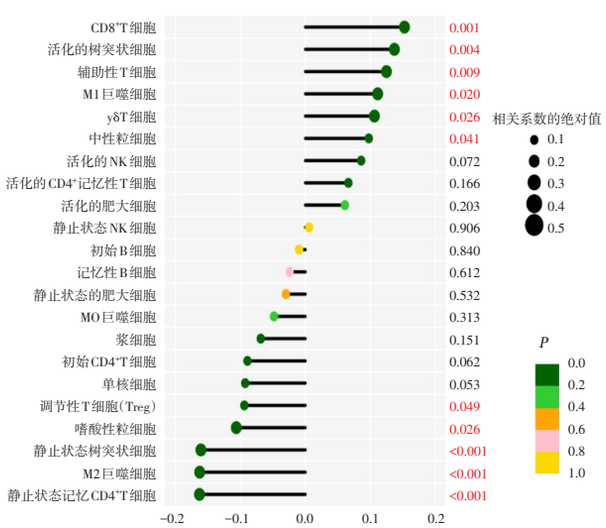

图3 *EIF5A*基因与肿瘤浸润免疫细胞的关系

Figure 3 The correlation between *EIF5A* and tumor-infiltrating immune cells

3 讨 论

MSI已经成为CRC诊断、治疗以及预后评价的最重要的临床特征之一。在肿瘤发生的研究领域中，MSI途径（约占15%）和染色体不稳定（chromosomal instability，CIN）（约占75%）途径成为散发性CRC发生的两大重要通路。尤其是Lynch综合征患者，几乎所有的患者都是经过MSI途径发生的[13]。与CIN CRC特征不同的是，MSI CRC主要发生在右半结肠，往往以黏液性和低分化腺癌为主的组织学类型出现。然而，MSI相关基因的研究仅限于微卫星位点和错配修复基因的改变。因此，本研究基于两种机器学习的算法和肿瘤免疫细胞浸润分析，最终在两个数据库中验证*EIF5A*可能是MSI的特征基因。

MSI-H在肿瘤病理诊断、肿瘤治疗和患者预后与MSI-L/MSS具有很大的差异，是现在CRC研究的热点之一。在肿瘤治疗中，II期MSI-H的CRC患者不适用5-氟尿嘧啶为主的化疗方案，而MSI-H的CRC患者对伊立替康等的化疗药物较为敏感[14]。在局部进展期低位直肠癌中，肠镜初诊活检组织中dMMR蛋白表型预示较好的新辅助放化疗疗效[15]。在肿瘤预后方面，有文献报道，MSI-H肿瘤预后优于MSI-L/MSS肿瘤[16]，尤其是在II期的CRC患者中[17]。本研究发现，GSE39582 CRC数据：MSI-H的患者预后要优于MSI-L/MSS患者。然而在TCGA的CRC患者数据中，MSI-H与MSI-L/MSS患者的生存时间差异无统计学意义。这可能与样品例数和种族有关。

在免疫治疗领域，MSI-H/dMMR患者已经公认为CRC免疫治疗的有效人群。MSI-H CRC患者在接受免疫检查点抑制剂后的客观缓解率为60%，疾病控制率为84%[18]。所有45例患者的12个月无疾病进展率为77%，12个月总体生存率为83%[18]。KEYNOTE-016研究[19]表明，62%（7/13）MSI-H的CRC患者预先接受过免疫检查点抑制剂治疗，并得到了客观缓解。KEYNOTE-164研究[20]表明，在接受一线治疗后的MSI-H的CRC患者再接受帕博利珠单抗治疗后，其客观缓解率为32%（中位随访时间为12.6个月），1年无进展生存率与总生存率分别为41%和76%。以上结论均一致表明：MSI成为CRC免疫治疗尤其是免疫检查点治疗的新型肿瘤标志物，因此，临床上关于MSI状态的辅助诊断和MSI影响免疫治疗的机制研究显得十分必要和迫切。

人工智能辅助诊断MSI方面，主要聚焦于病理切片信息[21]，病理多组学数据[22]，基因突变数据[23]等。在研究MSI状态对肿瘤免疫治疗的影响方面，Lin等[24]发现，与MSS/MSI-L型相比，MSI-H具有更多的免疫细胞浸润、更高的免疫相关基因表达和更高的免疫原性。此外，在肿瘤突变负荷

（tumor mutation burden，TMB）方面，与 MSS/MSI-L CRC（TMB<8 个突变/10^6 个 DNA 碱基）相比，MSI-H 具有更高的 TMB（>12 个突变/10^6 个 DNA 碱基）[25]。本研究通过 2 个独立的数据集（TCGA-COAD，GSE39582）层层筛选验证，使用机器学习的方式，最终确定了 *EIF5A* 基因为 MSI-H 的特征基因。在肿瘤免疫细胞浸润结果中，我们发现 *EIF5A* 基因表达水平与活化的树突状细胞，辅助性 T 细胞和 M1 巨噬细胞有关，这与 MSI-H CRC 拥有更高的活化淋巴细胞结果一致。*EIF5A* 是一个翻译起始因子，受羟腐胺赖氨酸作用调节。最新的研究数据表明，羟腐胺赖氨酸化的 *EIF5A* 能够调节如自噬[26]、衰老、多胺稳态[27]、能量代谢[28]等一系列关键的细胞进程，并在癌症[29]中起重要作用。Coni 等[30]发现：羟腐胺赖氨酸化的 *EIF5A* 可通过直接调节特定暂停状态下的 Myc 生物合成来促进 CRC 细胞的生长；而抑制 *EIF5A* 的羟腐胺赖氨酸化作用，可以抑制 CRC 细胞的生长。在具有家族性遗传性息肉病的小鼠模型中，阻断 *EIF5A* 羟腐胺赖氨酸化后腺瘤的抑制效果更佳明显。此外，文献[31]报道，聚腺苷二磷酸核糖水解酶（*PARG*）分子可以促进 Myc-MMR 轴，从而促进肿瘤的进展，同时也可以作为肿瘤免疫治疗的生物学标志物。虽然 *EIF5A* 与 MSI 患者的关系，以及 *EIF5A* 与免疫细胞浸润的关系尚未报道。本研究提出猜想，*EIF5A* 可能通过促进 Myc 的表达和延伸，从而促进 dMMR 的发生。靶向抑制 *EIF5A*（阻断其羟腐胺赖氨酸化作用），不仅可以作为 CRC 的潜在治疗方式，而且 *EIF5A* 的羟腐胺赖氨酸化有望成为 MSI 诊断和免疫检查点治疗的生物学标志物。

本研究仍然存在以下几点不足：首先，训练集和验证集来自美国和法国人群，其验证存在种族差异，而且还缺乏国内多中心的测序的验证结果。其次，关于 *EIF5A* 羟腐胺赖氨酸化–Myc-MMR 轴没有进行细胞验证，后续需要在基础实验中得以验证。

本研究基于 CRC 多个测序数据，首次发掘出 MSI 的特征基因 *EIF5A*，并发现其对 MSI 的诊断具有较高的准确度和效能，该基因有望成为 MSI 领域新的研究分子，为以后相关的功能机制研究提供线索和依据。

利益冲突：所有作者均声明不存在利益冲突。

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[J]. CA Cancer J Clin, 2018, 68(6):394-424. doi: 10.3322/caac.21492.

[2] Chen WQ, Zheng RS, Baade PD, et al. Cancer statistics in China, 2015[J]. CA Cancer J Clin, 2016, 66(2): 115-132. doi: 10.3322/caac.21338.

[3] Kawakami H, Zaanan A, Sinicrope FA. Microsatellite instability testing and its role in the management of colorectal cancer[J]. Curr Treat Options Oncol, 2015, 16(7):30. doi: 10.1007/s11864-015-2.

[4] 唐伟森, 廖明媚, 屈展, 等. 结直肠癌肿瘤组织 PMS2 蛋白表达状态与其临床病理特征的关系[J]. 中国普通外科杂志, 2019, 28(10):1297-1301. doi: 10.7659/j.issn.1005-6947.2019.10.019.
Tang WS, Liao MM, Qu Z, et al. Expression status of PMS2 protein in colorectal cancer tumor tissue and the relationship with its clinicopathological characteristics[J]. Chinese Journal of General Surgery, 2019, 28(10): 1297-1301. doi: 10.7659/j.issn.1005-6947.2019.10.019.

[5] Bartley AN, Luthra R, Saraiya DS, et al. Identification of cancer patients with Lynch syndrome: clinically significant discordances and problems in tissue-based mismatch repair testing[J]. Cancer Prev Res (Phila), 2012, 5(2): 320-327. doi: 10.1158/1940-6207.CAPR-11-0288.

[6] Yi M, Jiao DC, Xu HX, et al. Biomarkers for predicting efficacy of PD-1/PD-L1 inhibitors[J]. Mol Cancer, 2018, 17(1): 129. doi: 10.1186/s12943-018-3.

[7] Le DT, Durham JN, Smith KN, et al. Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade[J]. Science, 2017, 357(6349):409-413. doi: 10.1126/science.aan6733.

[8] Jung DH, Park HJ, Jang HH, et al. Clinical impact of PD-L1 expression for survival in curatively resected colon cancer[J]. Cancer Invest, 2020, 38(7): 406-414. doi: 10.1080/07357907.2020.1793349.

[9] Jiang YH, Yang M, Wang SH, et al. Emerging role of deep learning-based artificial intelligence in tumor pathology[J]. Cancer Commun (Lond), 2020, 40(4):154-166. doi: 10.1002/cac2.12012.

[10] Huang SG, Yang J, Fong S, et al. Artificial intelligence in cancer diagnosis and prognosis: opportunities and challenges[J]. Cancer Lett, 2020, 471:61-71. doi: 10.1016/j.canlet.2019.12.007.

[11] Cammarota G, Ianiro G, Ahern A, et al. Gut microbiome, big data and machine learning to promote precision medicine for cancer[J]. Nat Rev Gastroenterol Hepatol, 2020, 17(10): 635-648. doi: 10.1038/s41575-020-3.

[12] Chen BB, Khodadoust MS, Liu CL, et al. Profiling tumor infiltrating immune cells with CIBERSORT[J]. Methods Mol Biol,

　

2018, 1711:243−259. doi: 10.1007/978−1_12.

[13] Alicia L, Preethi S, Yelena K, et al. Microsatellite instability is associated with the presence of lynch syndrome pan-cancer[J]. J Clin Oncol Off J Am Soc Clin Oncol, 2019, 37(4):286−295. doi: 10.1200/JCO.18.00283.

[14] Bertagnolli MM, Niedzwiecki D, Compton CC, et al. Microsatellite instability predicts improved response to adjuvant therapy with irinotecan, fluorouracil, and leucovorin in stage III colon cancer: cancer and Leukemia Group B Protocol 89803[J]. J Clin Oncol, 2009, 27(11):1814−1821. doi: 10.1200/JCO.2008.18.2071.

[15] 程康文, 李佳, 王贵和, 等. 错配修复蛋白在直肠癌中的表达及其对新辅助放化疗敏感性的预测价值[J]. 中国普通外科杂志, 2020, 29(10): 1178−1186. doi: 10.7659/j. issn. 1005−6947.2020.10.004.
Cheng KW, Li J, Wang GH, et al. Expression of mismatch repair proteins in rectal cancer and its predictive value for sensitivity of neoadjuvant chemoradiotherapy[J]. Chinese Journal of General Surgery, 2020, 29(10): 1178−1186. doi: 10.7659/j. issn. 1005−6947.2020.10.004.

[16] Ma HY, Brosens LAA, Offerhaus GJA, et al. Pathology and genetics of hereditary colorectal cancer[J]. Pathology, 2018, 50(1): 49−59. doi: 10.1016/j.pathol.2017.09.004.

[17] Merok MA, Ahlquist T, Røyrvik EC, et al. Microsatellite instability has a positive prognostic impact on stage II colorectal cancer after complete resection: results from a large, consecutive Norwegian series[J]. Ann Oncol, 2013, 24(5):1274−1282. doi: 10.1093/annonc/mds614.

[18] Lenz HJJ, Cutsem EV, Limon ML, et al. Durable clinical benefit with nivolumab (NIVO) plus low-dose ipilimumab (IPI) as first-line therapy in microsatellite instability-high/mismatch repair deficient (MSI-H/dMMR) metastatic colorectal cancer (mCRC)[J]. Ann Oncol, 2018, 29:viii714. doi: 10.1093/annonc/mdy424.019.

[19] Le DT, Uram JN, Wang H, et al. PD-1 blockade in tumors with mismatch-repair deficiency[J]. N Engl J Med, 2015, 372(26):2509−2520. doi: 10.1056/NEJMoa1500596.

[20] Le DT, Kim TW, van Cutsem E, et al. Phase II open-label study of pembrolizumab in treatment-refractory, microsatellite instability-high/mismatch repair-deficient metastatic colorectal cancer: KEYNOTE-164[J]. J Clin Oncol, 2020, 38(1):11−19. doi: 10.1200/JCO.19.02107.

[21] Hildebrand LA, Pierce CJ, Dennis M, et al. Artificial intelligence for histology-based detection of microsatellite instability and prediction of response to immunotherapy in colorectal cancer[J]. Cancers (Basel), 2021, 13(3):391. doi: 10.3390/cancers13030391.

[22] Cao R, Yang F, Ma SC, et al. Development and interpretation of a pathomics-based model for the prediction of microsatellite instability in Colorectal Cancer[J]. Theranostics, 2020, 10(24): 11080−11091. doi: 10.7150/thno.49864.

[23] Krause J, Grabsch HI, Kloor M, et al. Deep learning detects genetic alterations in cancer histology generated by adversarial networks[J]. J Pathol, 2021, 254(1):70−79. doi: 10.1002/path.5638.

[24] Lin AQ, Zhang J, Luo P. Crosstalk between the MSI status and tumor microenvironment in colorectal cancer[J]. Front Immunol, 2020, 11:2039. doi: 10.3389/fimmu.2020.02039.

[25] Ganesh K, Stadler ZK, Cercek A, et al. Immunotherapy in colorectal cancer: rationale, challenges and potential[J]. Nat Rev Gastroenterol Hepatol, 2019, 16(6):361−375. doi: 10.1038/s41575−019-x.

[26] Zhang HL, Alsaleh G, Feltham J, et al. Polyamines control eIF5A hypusination, TFEB translation, and autophagy to reverse B cell senescence[J]. Mol Cell, 2019, 76(1): 110−125. doi: 10.1016/j.molcel.2019.08.005.

[27] Martella M, Catalanotto C, Talora C, et al. Inhibition of eukaryotic translation initiation factor 5A (eIF5A) hypusination suppress p53 translation and alters the association of eIF5A to the ribosomes[J]. Int J Mol Sci, 2020, 21(13):4583. doi: 10.3390/ijms21134583.

[28] Pelechano V, Alepuz P. eIF5A facilitates translation termination globally and promotes the elongation of many non polyproline-specific tripeptide sequences[J]. Nucleic Acids Res, 2017, 45(12): 7326−7338. doi: 10.1093/nar/gkx479.

[29] Zhang J, Li X, Liu XR, et al. EIF5A1 promotes epithelial ovarian cancer proliferation and progression[J]. Biomedecine Pharmacother, 2018, 100: 168−175. doi: 10.1016/j.biopha.2018.02.016.

[30] Coni S, Serrao SM, Yurtsever ZN, et al. Blockade of EIF5A hypusination limits colorectal cancer growth by inhibiting MYC elongation[J]. Cell Death Dis, 2020, 11(12): 1045. doi: 10.1038/s41419−020−6.

[31] Yu MC, Chen Z, Zhou Q, et al. PARG inhibition limits HCC progression and potentiates the efficacy of immune checkpoint therapy[J]. J Hepatol, 2022, 77(1): 140−151. doi: 10.1016/j.jhep.2022.01.026. [Online ahead of print]

（本文编辑　姜晖）

本文引用格式：李秀勤, 韩腾辉, 王帅, 等. 基于机器学习的结直肠癌微卫星不稳定基因挖掘及其应用价值分析[J]. 中国普通外科杂志, 2022, 31(10):1355−1362. doi：10.7659/j.issn.1005−6947.2022.10.011

***Cite this article as*:** Li XQ, Han TH, Wang S, et al. Mining of genes involved in microsatellite instability in colorectal cancer through machine learning and evaluation of their application values[J]. Chin J Gen Surg, 2022, 31(10): 1355−1362. doi：10.7659/j. issn. 1005−6947.2022.10.011